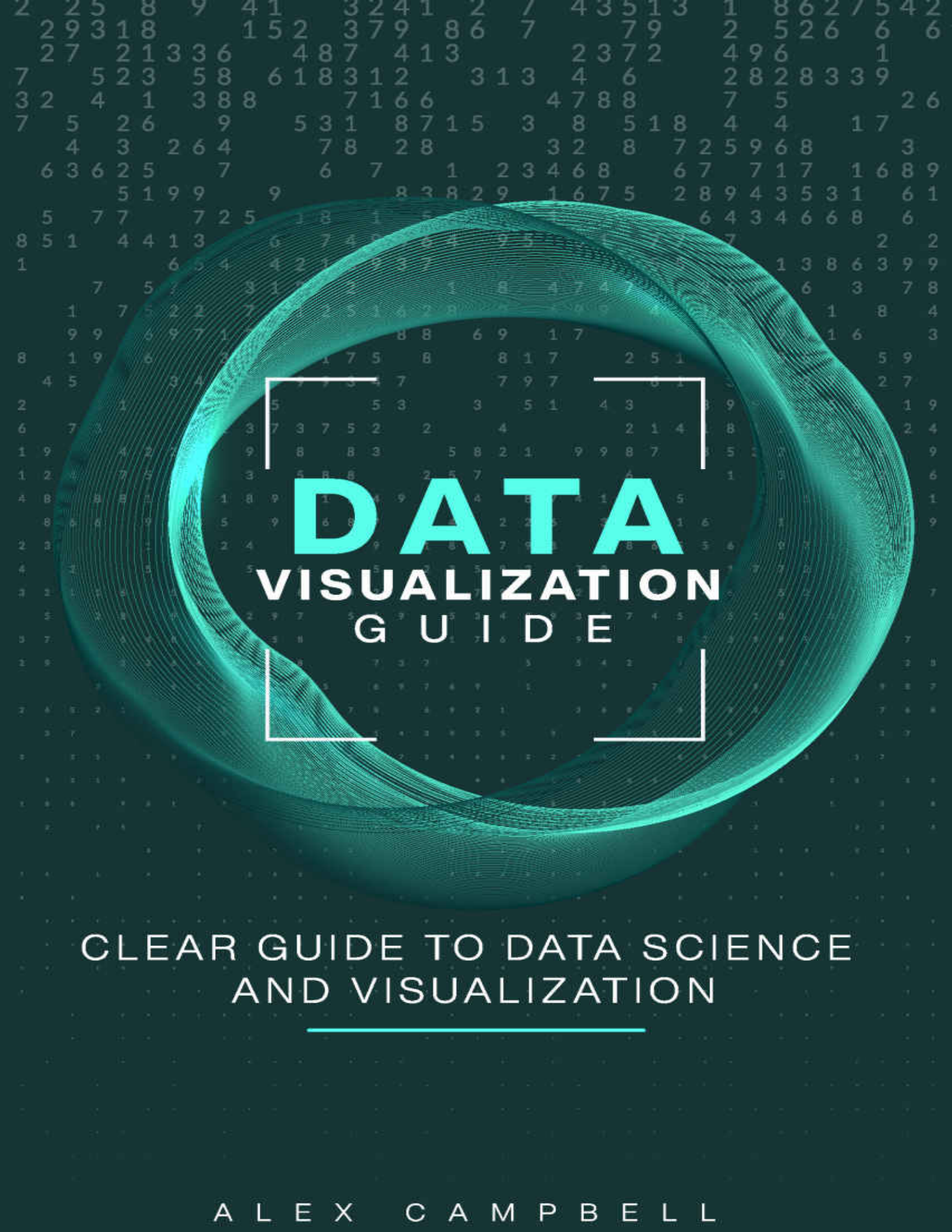


CLEAR GUIDE TO DATA SCIENCE AND VISUALIZATION

ALEX CAMPBELL

Data Visualization Guide

Clear Guide to Data Science and Visualization

Alex Campbell

OceanofPDF.com

© Copyright 2021 - All rights reserved. Alex Campbell.

The contents of this book may not be reproduced, duplicated or transmitted without direct written permission from the author.

Under no circumstances will any legal responsibility or blame be held against the publisher for any reparation, damages, or monetary loss due to the information herein, either directly or indirectly.

Legal Notice:

This book is copyright protected. This is only for personal use. You cannot amend, distribute, sell, use, quote or paraphrase any part or the content within this book without the consent of the author.

Disclaimer Notice:

Please note the information contained within this document is for educational and entertainment purposes only. Every attempt has been made to provide accurate, up to date and reliable complete information. No warranties of any kind are expressed or implied. Readers acknowledge that the author is not engaging in the rendering of legal, financial, medical or professional advice. The content of this book has been derived from various sources. Please consult a licensed professional before attempting any techniques outlined in this book.

By reading this document, the reader agrees that under no circumstances is the author responsible for any losses, direct or indirect, which are incurred as a result of the use of information contained within this document, including, but not limited to, —errors, omissions, or inaccuracies.

OceanofPDF.com

Table of Contents

Data Visualization Guide

BOOK 1 - Data Visualization Guide: Clear Introduction to Data Mining, Analysis, and Visualization.

INTRODUCTION

CHAPTER ONE: INTRODUCTION TO DATA MINING

Data Types Used in Mining

- **Relational Databases**
- **Data Warehouses**
- **Data Repositories**
- **Object-Relational Database**
- **Transactional Databases**

Pros and Cons of Data Mining

- **Pros**
- **Cons**

Applications of Data Mining

- **Healthcare**
- **Market Basket Analysis**
- **Education**
- **Manufacturing Engineering**
- **Customer Relationship Management (CRM)**
- **Fraud Detection**
- **Lie Detection**

OceanofPDF.com

Financial Banking

Challenges

Noisy and Incomplete Data
Data Distribution
Complex Data
Performance
Data Security and Privacy
Data Visualization

CHAPTER TWO: DATA MINING ARCHITECTURE

Data Source

Different Processes

Data Warehouse Server or Database

Data Mining Engine

Pattern Evaluation Module

Graphical User Interface (GUI)

Knowledge Base

CHAPTER THREE: DATA MINING TECHNIQUES

Classification

Clustering

Regression

Association Rules

Lift
Support
Confidence

Outer Detection

OceanofPDF.com

Sequential Patterns

Prediction

CHAPTER FOUR: DATA MINING TOOLS

Orange Data Mining
- **Why Should You Use Orange?**

SAS Data Mining

DataMelt Data Mining

Rattle

Rapid Miner

CHAPTER FIVE: INTRODUCTION TO DATA ANALYSIS

Why Use Data Analysis?

Data Analysis Tools

Data Analysis Types
- **Text Analysis**
- **Statistical Analysis**
- **Descriptive Analysis**
- **Diagnostic Analysis**
- **Predictive Analysis**
- **Prescriptive Analysis**

Data Analysis Process
- **Data Requirement Gathering**
- **Data Collection**
- **Data Cleaning**
- **Data Analysis**
- **Data Interpretation**
- **Data Visualization**

CHAPTER SIX: MANIPULATION OF DATA IN PYTHON

OceanofPDF.com

NumPy
- **Create Arrays**
- **Array Slicing**

Let's Start with Pandas
- **How to Summarize Data Using Pandas**

How to Work with Noise or Duplicate Records
- **Categorization of Data**
- **How to Solve Problems Using Pandas**
- **How to Clean Data Sets**
- **Using Pre-Defined Conditions**

CHAPTER SEVEN: EXPLORING THE DATA SET

CHAPTER EIGHT: HOW TO SUMMARIZE DATA WITH PYTHON

Obtain Information About the Data Set
- **Info()**
- **Describe()**
- **Value_counts()**
- **nunique()**
- **Sun()**
- **count()**
- **Minimum, Maximum, Mean, and Median**

CHAPTER NINE: STEPS TO BUILD DATA ANALYSIS MODELS IN PYTHON
- **Step One: Descriptive Analysis and Data Exploration**
- **Step Two: Remove Missing Values in the Data Set**
- **Step Three: Data Modeling**
- **Step Four: Estimate the Model's Performance**

CHAPTER TEN: HOW TO BUILD THE MODEL
- **Step One: Import Libraries Required to Develop the Model and Perform the Analysis. Read the Information and Split the Data into Training and Testing Data Sets.**
- **Step Two: Summarize the Data Set into Columns**
- **Step Three: Identification of Variables**

Step Four: Identify the Missing Values in the Data Set and Create a Flag

OceanofPDF.com

Step Five: Remove Missing Values
Step Six: Create Label Encoders for Categorical Variables
Step Seven: Remove Missing Values by Substituting the Dummy Variables
Step Eight: Check the Model's Performance and Make Predictions Based on the Analysis Results

CHAPTER ELEVEN: DATA VISUALIZATION

Know Your Audience

Set Your Goals

Choosing the Right Charts to Represent Data
Number Charts
Maps
Pie Charts
Gauge Charts

CONCLUSION

REFERENCES

BOOK 2 - DATA SCIENCE FOR BEGINNERS: COMPREHENSIVE GUIDE TO MOST IMPORTANT BASICS IN DATA SCIENCE.

INTRODUCTION

CHAPTER ONE: INTRODUCTION TO DATA SCIENCE

Introduction to Data Science
Predictive Casual Analytics
Prescriptive Analytics
Machine Learning

Why Use Data Science?
Customer Service

OceanofPDF.com

Self-Driven Cars
Predictions

Who is a Data Scientist?
Functions Performed by Data Scientists

Differences Between Data Science and Business Intelligence

Lifecycle
Phase One: Discovery
Phase Two: Data Preparation
Phase Three: Plan the Model
Phase Four: Build the Model
Phase Five: Operate the Model
Phase Six: Communicate the Results

CHAPTER TWO: PROS AND CONS OF DATA SCIENCE

Pros
Fastest Growing Field
Abundance of Roles
A Versatile Field
Makes Data Easier to Use
A Prestigious Career
Remove Redundancy
Products Become Smarter
Save Lives
Aid in Personal Growth

Cons
Data Science is an Ambiguous Term
Impossible to Master Data Science
Requires a Lot of Domain Knowledge
Unexpected Results
Lack of Data Privacy

CHAPTER THREE: STATISTICS FOR DATA SCIENCE

Descriptive Statistics
Normal Distribution

OceanofPDF.com

Variability
Kurtosis and Skewness

Inferential Statistics
Central Limit Theorem
Hypothesis Testing
ANOVA

CHAPTER FOUR: SKILLS REQUIRED FOR DATA SCIENCE

Technical Skills
Understanding Data
Algorithms
Programming
Analytical Tools
Using Unstructured Data

Non-Technical Skills
Strong Business Acumen
Communication Skills
Intuition

CHAPTER FIVE: TOOLS REQUIRED FOR DATA SCIENCE

R Programming Language

Python Programming Language

Hadoop

Structured Query Language (SQL)

SAS

Apache Spark

D3.js

CHAPTER SIX: APPLICATION OF DATA SCIENCE

OceanofPDF.com

Risk and Fraud Detection

Healthcare

Genetics and Genomics

Drug Development

Internet Search

Website Recommendations

Advanced Image Recognition

Virtual Assistance

Speech Recognition

Planning Routes for Airplanes

Gaming

Augmented Reality

CHAPTER SEVEN: ALGORITHMS AND STATISTICS

Naïve Bayes

Linear Regression

Logistic Regression

Neural Networks

- **K-Means Clustering**
- **Decision Tree**
- **Some Final Thoughts**

Applying Math to Data Science Models

- **Deriving Insights from Statistical Methods**
- **Some Essentials for Data Science**

OceanofPDF.com

CHAPTER EIGHT: DATA SCIENCE IN A SECTOR

Communicate the Insights Captured Through Data

Leverage Cloud-Based Solutions

CHAPTER NINE: DATA CLEANING USING R

Step One: Exploratory Analysis of Data

Step Two: Visual Exploratory Analysis

Histogram

Box Plot

Step Three: Correct Errors

CHAPTER TEN: DEVELOP A PREDICTIVE MODEL USING R

CHAPTER ELEVEN: CREATE DATA VISUALIZATION TO COMMUNICATE THE INSIGHT

CHAPTER TWELVE: DOCUMENT YOUR INSIGHTS FROM THE DATA

CONCLUSION

REFERENCES

OceanofPDF.com

Data Visualization Guide

Clear Introduction to Data Mining, Analysis, and Visualization

Alex Campbell

OceanofPDF.com

Introduction

Most organizations and businesses collect large volumes of data from various sectors and departments. This data is often unformatted, so you will need to find a way to process and clean it. Businesses can then use this information to make informed business decisions. They use data analysis and mining to interpret the data and collect the necessary information from the data set. These processes play an important role in any business. You can also use this type of analysis in your personal life. Data mining and analysis can be used to help you save money. Only when businesses know how to work with data can they know where they should reinvest the money and increase their revenue.

If you are new to the world of data, this book can be your guide. You can use the information to help you learn the basics of data mining and analysis. The book will also shed some light on the processes you can use to clean the data set, various processes and techniques you can use to mine and analyze information, and it will explain to you how you can visualize the data and why it's important to represent data using graphs and other visuals.

Within these pages you will find information about the different techniques and algorithms used in data analysis, as well as provide you with different libraries you can use to manipulate and clean data sets. Most data analysis and mining algorithms are built using Python, and thus we will use the libraries and functions from Python in the book. You will also find a section including information about the process used to develop a model.

OceanofPDF.com

Before you work on developing different analysis techniques, you need to make sure you have the business problem or query in mind. It is important to bear in mind that any analysis you perform should be based on a business question. You need to make sure there is a foundation upon which you develop the model. Otherwise, the effort you put in will be unusable. Make sure you have all the details about why you are developing a model or collecting information before you put in the effort.

OceanofPDF.com

Chapter One: Introduction to Data Mining

I am sure you may have heard many people talk about data mining and how essential it is. But what is data mining? As the name suggests, data mining is the process of identifying and extracting hidden patterns, variables, and trends within any data set collected for your analysis. In simple words, the process of looking at data to identify any hidden patterns and trends of information that can be used to categorize the data into useful analysis is termed data mining or knowledge discovery of data (KDD). You can use data mining to convert raw data or information into data, which businesses can use.

It is important to remember that organizations often collect and assemble data from data warehouses. They use different data mining algorithms and efficient analysis algorithms to make informed decisions about their business. Through data mining, businesses can go through large volumes of data to identify patterns and trends, which would not be possible through simple analysis algorithms. We use complex statistical and mathematical algorithms to evaluate data segments and calculate a future event's probability. Organizations use data mining to extract the required information from large databases or sets to answer different business questions or problems.

Data mining and science are similar to each other, and in specific situations, these processes are carried out by one individual.

OceanofPDF.com

There is always an objective for these processes to be performed. Data science and data mining processes include web mining, text mining, video and audio mining, social media mining, and pictorial data mining. This can be done with ease through different software.

Companies should outsource data mining processes since they have a lower operation cost. Some firms also use technology to collect various forms of data that cannot be located manually. You can find large volumes of data on different platforms, but there is very little knowledge that can be accessed from this data.

Every organization finds it difficult to analyze the various information collected to extract the information needed to solve any problem or make informed business decisions. There are numerous techniques and instruments available to mine information from various sources to obtain necessary insights.

Data Types Used in Mining

You can perform data mining on the following data types:

Relational Databases

Every database is organized in the form of records, tables, and columns. A relational database is one where you can collect information from different tables or data sets and organize it in the form of columns, records, and tables.

OceanofPDF.com

You can access this data easily without having to worry about individual data sets. The information is conveyed and shared through tables that increase the ease of organization, reporting, and searchability.

Data Warehouses

Every business collects data from different sources to obtain information to help them make well-informed decisions. They can do this easily using the process of data warehousing. These large volumes of data come from different sources, such as finance and marketing. The extracted information is then used for the purpose of analysis, which helps businesses make the right organizational decisions. A data warehouse has been designed to analyze data and not to process transactions.

Data Repositories

A data repository refers to the location where the organization can store data. Most IT professionals use this term to refer to the setup of data and its collection in the firm. For instance, they term a group of databases where different kinds of data are stored.

Object-Relational Database

An object-relational database is a combination of a relational database and an object-oriented database model. This database uses inheritance, classes, objects, etc. This database aims to close the gap between an object-oriented and relational database model by using different programming languages, such as C#, C++, and Java.

OceanofPDF.com

Transactional Databases

A database management system, also known as a transactional database, can reverse any transacti0n made in the database if it was not performed in the right way. This is a unique capability, which was defined a while back in the form of a relational database. These are the only databases that support such activities.

Pros and Cons of Data Mining

Pros

- Data mining techniques enable organizations to obtain information and trends from the data set. They can use this information to make informed decisions for the organization
- Through data, mining organizations can make the necessary modifications in production and operation
- Data mining is not as expensive as other forms of statistical data analysis
- Businesses can discover hidden trends and patterns in the data set and calculate the probabilities of the occurrence of specific trends
- Since data mining is an easy and quick process, it does not take too long to introduce it onto a new or existing platform. Data mining processes and algorithms can be used to analyze large volumes of data in a short span.

OceanofPDF.com

Cons

- Data privacy and security is one of the major concerns of data mining. Organizations can sell their customers' data to other organizations for money. American Express has sold the purchases of their customers to other organizations for money
- Most data mining software uses extensive and difficult algorithms to operate, and any user working on these algorithms needs the required training to work on those models
- Different models work in different ways because of the different algorithms and concepts used in those models. Therefore, it is important to choose the right data mining model
- Some data mining techniques do not produce accurate results, and this can lead to severe repercussions

Applications of Data Mining

Most organizations with intense demands from consumers use data mining. Some of these organizations include communication, retail, marketing company, finance, etc. They use it for the following:

1. To identify customer preferences
2. Understand how customers can be satisfied
3. Assess the impact of various processes on sales
4. The positioning of products in the organization
5. Assess how to improve profits

OceanofPDF.com

Retailers can use data mining to develop promotions and products to attract their customers. This section covers some areas where data mining is used widely.

Healthcare

Data mining can improve different aspects of the health system since it uses both analytics and data to obtain better insights from the data sets. The healthcare industry can use this information to identify the right services to improve health care services and reduce costs. Most businesses also use data mining approaches, such as data visualization, machine learning, soft computing, statistics, and multi-dimensional databases, to analyze different data sets and forecast the patients in different categories. These data mining processes enable the healthcare industry to ensure patients obtain the necessary intensive care at the right time and place. Data mining also enables insurers to identify any abuse and fraud.

Market Basket Analysis

This form of analysis is based on different hypotheses. If you purchase specific products, you will probably buy another product from the same product group. This form of analysis makes it easier for any retailer to identify any customer's purchase behavior in the customer group. The retailer can also use this information to understand what a buyer or customer wants or needs, making it easier for them to alter the store's layout.

OceanofPDF.com

You can also make a comparison between different stores that make it easier for you to differentiate between different customer groups.

Education

The use of data mining in education is relatively new, and the objective of using data mining in this industry is to explore knowledge from large volumes of data from educational environments. Educational data mining (EDM) can be used to understand the future behavior of a student by studying the impact of various educational systems and support on the student. Educational organizations use data mining to make the right decisions to help students improve. They also use data mining to predict a student's results. Educational institutions can use this information to identify what a student should be taught. They can also use this information to define how to teach students.

Manufacturing Engineering

Every manufacturing company must know what the customers want, and this knowledge is their asset. You can use various data mining tools to help you identify any hidden patterns and trends in various manufacturing processes. You can also use data mining to develop the right company designs and obtain any correlation between product portfolios and architecture. You can incorporate different customer requirements to develop a model that caters to both the business and customer needs. This information can then be used to forecast product development, cost, and delivery dates, among other criteria.

OceanofPDF.com

Customer Relationship Management (CRM)

The objective of CRM is to obtain and maintain customers, thereby enabling businesses to develop customer-oriented strategies and enhance customer loyalty. If you want to improve your relationship with customers, you need to collect the right information and analyze it accurately. When you use the right data mining technologies, you can use the data collected to analyze and identify methods to improve customer relationships.

Fraud Detection

Have you ever loaned someone money, and they ghosted you immediately after? That is an example of fraud, but this is only on a small scale. Banks and other financial institutions lose close to a billion dollars each year because of fraudulent customers. Traditional fraud detection methods are sophisticated and time-consuming. Data mining techniques and methods use different statistical and mathematical algorithms to identify hidden and meaningful data set patterns.

A fraud detection system should be used to protect all the information in the data set while protecting each user's data. Supervised data mining models have a collection of training or sample records using which the model can classify some customers as frauds. You can construct a model using this information. The objective of this model is to identify if there are fraudulent claimants and documents or not.

OceanofPDF.com

Lie Detection

It is not difficult to apprehend criminals, but it is extremely difficult to bring the truth out of them. This is a very challenging task. Many police departments and law enforcement agencies now use data mining techniques to minor any communication between suspected terrorists, investigate prior offenses, etc.

The data mining algorithms used for this also include text mining. In this process, the algorithm goes through various text files and data to identify hidden patterns in the data set. The data used in this format is often unstructured. These algorithms compare the current output against previous outputs to develop a lie detection model.

Financial Banking

Banks have now taken a turn and have started digitizing all the transactions and information stored by customers. Using data mining algorithms and techniques, bankers can solve various business-related issues and problems.

They can use these models to identify various trends, correlations, and patterns in the data collected. Bankers can use these methods when they work with large volumes of data. It is easier for managers and experts to use these data and correlations to better acquire, target, segment, maintain and retain various customer profiles.

OceanofPDF.com

Challenges

Data mining is an important process and extremely powerful. There are, however, many challenges you may face when you implement or execute these algorithms. These challenges are related to data, performance, techniques, and methods used in data mining. The data mining process becomes effective only when you identify the problems or challenges and resolve them effectively.

Noisy and Incomplete Data

As mentioned earlier, the process of extracting useful information and trends from large volumes of data is termed data mining. It is important to remember that data collected in real-time is incomplete, noisy, and heterogeneous. It is difficult to determine if the data collected is reliable or accurate.

These are some problems that occur because of human errors or inaccurate measuring instruments. Let us assume you run a retail chain. Your job is to collect the number of every customer who spends more than $1000 at your store. When you identify such a customer, you send a notification to the accounting person, who then enters the information.

The accounting person can enter the incorrect number in the data set, which will lead to incorrect data. Some customers may also enter the wrong number in a hurry or for other reasons. Other customers may not want to enter their number for privacy reasons. These situations make the data mining process challenging.

OceanofPDF.com

Data Distribution

Real-world data is stored on numerous platforms in a computing environment distributed across different platforms. The data can be stored on the Internet, individual systems, or in a database. This makes it difficult to shift the data from these sources into a central data repository due to different technical and organizational concerns. For instance, some regional offices may have their data stored on their servers to store the data. It is impossible to store the data from different regional offices on one server if you think about it. Therefore, if you want to mine data, you need to develop the necessary algorithms and tools which make it easier to mine large volumes of data.

Complex Data

Businesses now collect data from different sources, and this data is heterogeneous in nature. It can include different multimedia data, such as video, audio, and images, and other complex data, such as time series, spatial data, and so on. It is difficult for anybody to manage this data and analyze it or extract useful information from it. New tools, methodologies, and technologies must be refined most times if you want to obtain the required information.

Performance

The performance of any data visualization model relies on the algorithm being used and its efficiency. The technique with which the model is developed also determines the performance of the model.

OceanofPDF.com

If the algorithm designed is not built correctly, the efficiency of the process is significantly affected.

Data Security and Privacy

Data mining can lead to a serious issue when it comes to data governance, privacy and security. Let us assume you are a retailer who analyzes a customer’s purchasing patterns. To do this, you need to collect all your customers' data, purchasing habits, preferences, and other details. You need to collect this information, and you may not require your customer’s permission to do this.

Data Visualization

Data visualization is an important process in data mining. This is the only way you or a business can visualize the different patterns and trends in the data set. Businesses and data scientists need to identify what the data and variables in the data set are trying to convey. It is also important to know what the data is trying to express to you. There are times when it is not easy to present the data in an easy-to-understand manner. Some input data points, or variables, may produce complicated outputs. Therefore, you need to identify efficient and accurate data visualization processes if you want to succeed.

OceanofPDF.com

Chapter Two: Data Mining Architecture

Now that you have a basic idea of data mining, let us learn more about the data mining architecture. Some significant components of data mining are the data mining engine, data source, the pattern evaluation module, knowledge base, data warehouse servers, and graphical user interface. Let us look at each of these components in further detail.

Data Source

Data can be sourced from the following:

- Database
- Data warehouse
- Internet
- Text files

If you want to obtain useful information from data mining, you need to collect large volumes of historical information. Most organizations store data in data warehouses or databases.

A data warehouse can include more than one database, a data repository, or text file spreadsheets. You can also use spreadsheets or text files as the source since they can contain some information.

OceanofPDF.com

Different Processes

Before the collected data or information is moved into a data warehouse or database, the information should be processed, selected, integrated, and cleaned. Since information is collected from numerous sources that store data in different formats, you cannot use it directly to perform any data mining operations. The results of the data mining process will be inaccurate and incomplete if you use unstructured data. Therefore, the first step in the process is to clean the data that you need to work with and then pass it onto the server. The process of cleaning the data is not as easy as one thinks. You can perform different kinds of operations on the data as part of the cleaning, integration, and selection.

Data Warehouse Server or Database

Once you select the data you want to use from different sources, you can clean it and pass it onto the data warehouse server or database. This is the source of the original data that you will process and use in the data mining process. The user uses the server, meaning you or the business, to retrieve the information relevant to the data mining request.

Data Mining Engine

This is a very important component of the data mining architecture since it contains different modules that can be used to perform various data mining tasks. These include:

- Classification

OceanofPDF.com

- Characterization
- Association
- Prediction
- Clustering
- Time-series analysis

In simple words, we can say that this engine is the root or base of the entire architecture. The engine comprises different software and instruments that can be used to obtain knowledge and insights from the mining process's data. You can also use the engine to learn more about any kind of data stored in the server.

Pattern Evaluation Module

This model is used in the data mining architecture to measure or investigate the pattern, followed by the variables based on a threshold value. This module works with the data mining engine to identify various patterns in the data set. The pattern evaluation module uses different stake measures that cooperate with various data mining modules in the engine to identify different patterns or trends in the data sets. This module uses a stake threshold to locate any hidden patterns and trends in the data set.

The pattern evaluation module can work with the mining module, but this is only dependent on the different techniques used in the data mining engine. If you want to develop efficient and accurate data mining models, you need to push the evaluation of this stake measure as much as possible into the mining process. This will ensure the model only looks at the different patterns in the data set.

OceanofPDF.com

Graphical User Interface (GUI)

The GUI is one of the data mining architecture modules that communicate between the user and the data mining system or module. This module helps users to efficiently and easily communicate with the system without worrying about how complex the process is. The GUI module works with the data mining system based on the user's task or query to display the required results.

Knowledge Base

This is the last module in the data set, which helps the entire data mining process. This module can be used to evaluate the stake measure used to identify hidden results and guide the search. The knowledge base module contains data from a user’s experience, user views, and other information, which helps in the data mining process. The knowledge base obtains inputs and information from the data mining engine to obtain reliable and accurate information. The knowledge base also interacts with the pattern assessment module to obtain inputs and also update the data stored in the module.

OceanofPDF.com

Chapter Three: Data Mining Techniques

Now, let us look at some data mining techniques, which can be incorporated into the data mining engine. These techniques allow you to identify hidden, valid, and unknown patterns and correlations in the large data sets. These techniques use different machine learning techniques, mathematical algorithms, and statistical models to answer different questions.

Some examples of such algorithms are neural networks, decision trees, classification, etc. Data mining predominantly uses prediction and analysis.

Data mining professionals use different methods to understand, process, and analyze data to obtain accurate conclusions from large volumes of data. The methods they use are dependent on various technologies and methods from the intersection of statistics, machine learning, and database management. So, what are the methods they use to obtain these results?

In most data mining projects, professionals have used different data mining techniques. They have also developed and used different modules and techniques, such as classification, association, prediction, clustering, regression, and sequential patterns. We will look at some of these in further detail in this chapter.

OceanofPDF.com

Classification

The classification technique is used to obtain relevant and important information about the metadata and data used in the mining process. Professionals use this technique to classify data points and variables into different classes. Some of these techniques can be classified into the following:

1. We can classify various data mining frameworks and techniques based on the source of data you are trying to mine. This process is based on the data you use or handle. For instance, you can classify data into time-series data, text data, multimedia, World Wide Web, spatial data, etc.
2. Data can be classified into different frameworks based on the database you use in your analysis. This type of classification is based on the type of model you are using. For instance, you can classify the data into the following categories: relational database, object-oriented database, transactional database, etc.
3. We can classify data into a framework based on the type of knowledge extracted from the data set. This form of classification is dependent on the type of information you have extracted from the data. You can also use the different functionalities used to perform this classification. Some frameworks used are clustering, classification, discrimination, characterization, etc.
4. Data can also be classified into a framework based on the different techniques used to perform data mining. This form of classification is based on the approach of analysis used to mine the data, such as machine learning, neural networks, visualization, database-oriented and data warehouse-oriented, genetic

OceanofPDF.com

algorithms, etc. This form of classification also takes into account how interactive the GUI is.

Clustering

Clustering is an algorithm used to divide information into groups of connected objects based on their characteristics. When you divide the data set into clusters, you may lose some details present in the data set, but you improve the data set. When it comes to data modeling, clustering is rooted in mathematics, numerical analysis, and statistics. If you look at data modeling in terms of machine learning, the clusters show hidden patterns in the data set.

The model looks for clusters in the data set using unsupervised machine learning. The subsequent framework developed will represent a concept of data. When you look at it from a practical point of view, this form of analysis plays an important role in various data mining applications, such as text mining, scientific data exploration, spatial database applications, Web analysis, CRM, computational biology, information retrieval, medical diagnostics, etc.

In simple words, clustering analysis is a form of data mining technique used to identify the data points in the data set, which share numerous similarities.

This technique will help you recognize various similarities and differences in the data set. Clustering is similar to classification, but in this technique, you group large chunks of data into groups based on the similarities.

OceanofPDF.com

Regression

Regression analysis is another form of data mining, which is used to analyze and identify the relationship between different variables based on the presence of another variable in the data set. This technique is used to define the probability of the occurrence of a specific variable in the data set. The process of regression is a form of modeling and planning used in different algorithms. For instance, you can use this technique to project a cost or expense based on various factors, such as consumer demand, competition, and availability. This technique will give you the exact relationship between the variables in the data set. Some forms of this technique are linear regression, multiple regression, logistic regression, etc.

Association Rules

Data mining's association technique is to define a link between various data points in the data set. Using this technique, data mining professionals can identify any hidden patterns or trends in the data set. An association rule is a conditional statement using the if-then format, and these rules help the professional identify the probability of interactions between different data points in large data sets. You can also identify correlations between different databases as well.

The association rule mining technique is used in different applications and is often used by retailers to identify correlations in medical data sets or retail data sets. The algorithm works differently on different data sets.

OceanofPDF.com

For instance, you can collect the data of all the items you purchased in the last few months and run some association rules on the items to see what you want to purchase together. Some measurements used are:

Lift

This measurement is used to define the accuracy of the probability of how often you have purchased a specific product. The formula used to do this: (confidence interval) / (item A) / (entire data set).

Support

This process is used to determine how often you purchase different items and compares that to the overall data collected. The formula used to do this is: (item C + item D) / (entire data set).

Confidence

This process measures the number of times you purchase a specific product when you purchase another product as well. To do this, you can use the following formula: (item C + item D) / (item D)

Outer Detection

The outer detection technique is used when you need to identify the patterns or data points in the data set that do not match the expected data set behavior or pattern.

OceanofPDF.com

This technique is often used in domains such as detection, intrusion, fraud detection, etc. Outer detection is also known as outlier mining or outlier analysis.

An outlier is any point in the data set that does not behave in the same way as the average set of data points in the dataset. Most data sets have outliers in them, and this should be expected.

This technique plays an important role in the field of data mining. This technique is used in different fields, such as debit or credit card fraud detection, detection of outliers in wireless sensor network data, network interruption identification, etc.

Sequential Patterns

A sequential pattern is another technique used in data mining, which specializes in the evaluation of sequential information in the data set.

This is one of the best ways to identify any hidden sequential patterns in the data set.

This technique uses different subsequences from a single sequence where the stake cannot be measured in terms of various criteria, such as occurrence frequency, length, etc. In simple words, this process of data mining allows users to recognize or discover different patterns, either small or big, in the transaction data over a certain period.

OceanofPDF.com

Prediction

This technique uses a combination of different data mining processes and techniques, such as clustering, trends, classification, etc. Prediction looks at historic data, in the form of instances and events, in the appropriate sequence to predict the future.

OceanofPDF.com

Chapter Four: Data Mining Tools

As mentioned earlier, data mining uses a set of techniques with specific algorithms, statistical analysis, database systems, and artificial intelligence to analyze and assess data from various data sources and from different perspectives. Most data mining tools can discover different groupings, trends, and patterns in a large data set. These tools also transform the data into refined information.

Some of these frameworks and techniques allow you to perform different actions and activities that are key for data mining analysis. You can perform different algorithms, such as classification and clustering, on your data set using these data mining tools and techniques. The techniques use a framework, which provides insights for the data and various phenomena represented by the data set. These frameworks are termed data mining tools.

This chapter covers some common data mining tools used in the industry.

Orange Data Mining

Orange is a suite which uses different machine learning and data mining software. It also supports data visualization and is a software component, which is written in Python. This application was developed by the faculty of information and computer science from Ljubljana

OceanofPDF.com

University, Slovenia. Since this is an application that uses software-based components, the application's components are termed widgets. The different widgets used in the application can be used for preprocessing and data visualization. Using these widgets, you can assess different algorithms for data mining and also use predictive modeling. These widgets have different functionalities, such as

- Data reading
- Displaying data in the form of tables
- Selection of certain features from the data set
- Comparison between different learning algorithms
- Training predictors
- Data visualization

Orange also provides an interactive interface which makes it easier for users to work with different analytical tools. It is easy to operate the applications.

Why Should You Use Orange?

If you have data you collect from different sources, it can be formatted and arranged quickly in this application. You can format the data, so it follows the required pattern and move the widgets around to improve the interface. Different users can use this application since it allows a user to make a smart decision in a short time.

You can do this by analyzing and comparing data. Orange is a great way to visualize open-source data.

OceanofPDF.com

It also enables users to evaluate different data sets. You can perform data mining using different programming languages, including Python and visual programming. You can perform different analyses on this platform.

The application also comes with machine learning components, text mining features, and add-ons for bioinformatics. It also comes with different data analytics features and comes with a python library.

You can run different Python scripts on a terminal window or use an integrated environment, such as PythonWin, PyCharmand, and pr shells like iPython. The application has a canvas interface on which you can place the widget. You can then create a workflow in the interface to analyze the data set.

The widget can also perform fundamental operations, such as showing a data table, reading the data, training predictors, selecting required features from the data set, visualizing data elements, comparing learning algorithms, etc. The application works on various Linux operating systems, Windows and Mac OS X. It also comes with classification and multiple regression algorithms.

This application also reads documents in their native or other formats. It uses different machine learning techniques to classify the data into categories or clusters to aid in supervised data mining. Classification algorithms use two types of objects, namely classifiers and learners. A learner is termed class-leveled data, and it uses the data to return the classifier. You can also use regression methods in orange, and these are similar to classification methods; both techniques are designed for supervised data mining, and they need class-level data.

OceanofPDF.com

The ensemble will continue to learn using a combination of predictions from the individual model and precision measures. The model you develop can come from using different learners or different training data on the same data sets.

You can also diversify learners by changing the parameter set used by the learners. In this application, you can create ensembles by wrapping them around different learners. These ensembles also act like other learners, but the results they return allow you to predict the results for any other data instance.

SAS Data Mining

The SAS institute developed SAS or the Statistical Analysis System, and it is used for data management and analytics. You can use SAS to mine data, manage information from different sources, change the data and analyze the data's statistics. If you are a non-technical user, you can also use the graphical user interface to communicate with the application. The SAS data miner analyzes large volumes of data that provide the user with accurate insights to make the right decisions. SAS uses distributed memory processing architecture, which can be scaled in different ways. This application can be used for data optimization, mining, and text mining.

DataMelt Data Mining

DataMelt, also known as DMelt, is a visualization and computation environment that offers users an interactive

OceanofPDF.com

structure that can be used for data visualization and analysis. This application was designed especially for data scientists, engineers, and students. This application uses a multi-platform utility and is written in Java programming language. You can run this application on different operating systems as long as they are compatible with a Java Virtual Machine (JVM). This application consists of mathematics and science libraries.

- **Mathematical Libraries**: The libraries used in this application are used for algorithms, random number generation, curve fitting, etc.
- **Scientific Libraries**: The libraries used are for drawing 3D or 2D plots.

DMelt can be used to analyze large volumes of data, statistical analysis, and data mining. This application is used in financial markets, engineering, and natural sciences.

Rattle

This tool or application is a tool that uses a graphic user interface (GUI). Rattle is developed using the R programming language. The application also exposes R's statistical power and offers different data mining features that can be used during the mining process. Rattle has a well-developed and comprehensive user interface and includes an integrated log tab, allowing users to produce the code to perform different GUI operations. You can use the application to produce data sets, and you can edit and view them. The application also allows you to review the code, use it for different purposes, and extend that code without any restrictions.

OceanofPDF.com

Rapid Miner

Most data mining professionals use rapid miner to perform predictive analytics. This tool was developed by a company named Rapid Miner. The code is written in Javascript language, and it offers users an integrated environment they can use to perform various operations apart from predictive analysis, such as deep learning, machine learning, and text mining.

Rapid miner can be used in different applications, such as commercial applications, education, research, training, machine learning, application development, and company applications. Rapid miner also provides users with a server on-site. It also allows users to use both private or public cloud infrastructure to store the data and perform operations on that data set. The base of this application is a client/server model. This application is relatively accurate when compared to other applications and tools and uses a template-based framework.

OceanofPDF.com

Chapter Five: Introduction to Data Analysis

Now that you have an idea of what data mining is, let us understand what data analysis is and the different processes used in data analysis in brief. We will look at these concepts in further detail later in the book.

Data analysis is the process of transforming, cleaning, and modeling any information collected to identify hidden patterns and information in the data set to make informed decisions.

Data analysis aims to extract useful information hidden in the data set and take the required decision based on the results of the analysis.

Why Use Data Analysis?

If you want to grow in life or improve your business, you need to perform some analysis on the data collected. If your business does not grow, you need to go back and acknowledge what mistakes you made and overcome those mistakes.

You also need to find a way to prevent these mistakes from happening again. If your business is growing, you need to look forward to making the necessary changes to your processes to make sure the businesses grow more. You need to analyze the information based on the business processes and data.

OceanofPDF.com

Data Analysis Tools

You can use different data analysis tools to manipulate and process data. These tools make it easier for you to analyze the correlations and relationships between different data sets. These tools also make it easier for you to identify any hidden insights or trends in the data set.

Data Analysis Types

The following are the different forms of data analysis.

Text Analysis

This form of analysis is also called data mining, and we have looked at this in great detail in the previous chapters.

Statistical Analysis

Using statistical analysis, you can define what happened in a certain event based on historical information. This form of analysis includes the following processes:

1. Collection
2. Processing of information
3. Analysis
4. Interpretation of the results
5. Presentation of the results
6. Data modeling

Statistical analysis allows you to analyze sample data. There are two forms of statistical analysis:

OceanofPDF.com

Descriptive Analysis

In this form of analysis, you look at the entire data set or only a portion of the summarized data set in the form of numerals. You can use this numerical information to calculate the values of central tendency.

nferential Analysis

In this analysis, you look at a sample from the entire data set. You can select different samples and perform the same processes to determine how the data set is set. This form of analysis also tells you how the data set is structured.

Diagnostic Analysis

Diagnostic analysis is used to determine how a certain event occurred. This type of analysis uses statistical models to identify any hidden patterns and insights from the data set. You can use diagnostic analysis to identify any new business process problems and see what caused that problem. You can also identify any similar patterns within the data set to see if there were any other problems with similar patterns. This form of analysis enables you to use prescriptions for any new problems.

Predictive Analysis

Predictive analysis is the use of historic data to determine what may happen in the future. The simplest example is where you decide what purchases you want to make.

OceanofPDF.com

Let us assume you love shopping and bought four dresses after dipping into your savings. Now, if your salary were to double the next year, you can probably buy eight dresses. This is an easy example, and not every analysis you do will be this easy. You need to think about the various circumstances when you perform this analysis since the prices of clothes can change over the next few months.

Predictive analysis can be used to predictions about future outcomes based on historical and current data. It is important to note that the results obtained are only forecasts. The accuracy of the model used is dependent on the information you have and how you can dig into it.

Prescriptive Analysis

The process of prescriptive analysis uses a combination of the insights and results of previous analyses and the action you want to take to solve a current decision or problem. Most companies are now data-driven, and they use this form of analysis since they need both descriptive and prescriptive analyses to improve data performance. Data professionals use these technologies and tools to analyze the data they have and derive the results.

Data Analysis Process

The process followed for data analysis is solely dependent on the information you gather and the application or tool you use to analyze and explore the data. You need to find patterns in the data set.

OceanofPDF.com

Based on the data and information you collect, you can make the necessary information to obtain the ultimate result or conclusion. The process followed is:

- Data Requirement Gathering
- Data Collection
- Data Cleaning
- Data Analysis
- Data Interpretation
- Data Visualization

Data Requirement Gathering

When it comes to data analysis, you need to determine why you want to perform this analysis. The objective of this step is to determine what the aim of your analysis is. You should decide what type of analysis you want to perform and how you want to perform this analysis. In this step, you should determine what you want to analyze and how you plan to measure or analyze this information. It is important to determine why you need to investigate and identify the measures you want to use to perform this analysis.

Data Collection

After you gather the requirement, you will obtain a clear idea of what data you have and what you need to measure. You will also know what to expect from your findings. It is now time for you to collect your data based on the requirements of your business. When you collect the data from the sources, you need to process and organize it before you analyze it.

OceanofPDF.com

Since you collect data from different sources, you need to maintain a log with the date of collection and information about the source.

Data Cleaning

The data you collect may be irrelevant for you or may not be useful for your analysis. Therefore, you need to clean it before you perform these processes. The data may contain white spaces, errors, and duplicate records, and thus it should be cleaned and free of errors. This should be done before you analyze the data because your analysis results are based on how well you clean the data.

Data Analysis

When you collect, process, and clean data, you can analyze it. When you manipulate data, you need to find a way to extract the information from the data set. If you do not find the necessary information, you need to collect more information from the data set.

During this phase of the process, you can use the tools, techniques, and software for data analysis, which will enable you to interpret, understand, analyze, and extract necessary conclusions based on the requirement.

Data Interpretation

When you analyze the data completely, it is time for you to interpret the results. When you have the results, you can either use a chart or table to display the analysis. You can use the results of the analysis to identify the best action you can take.

OceanofPDF.com

Data Visualization

Most people use data visualization regularly, and they use often appear in the form of graphs and charts. In simple words, when you show data in the form of a graph, it is easier for the brain to understand the information and process it. Data visualization is used to identify any hidden facts, trends, and correlations in the data set. When you observe the relationships or correlations between the data points, you can obtain meaningful and valuable information.

OceanofPDF.com

Chapter Six: Manipulation of Data in Python

Data processing and cleaning is an important aspect of data analysis. This chapter sheds some light on the different ways you can use the Pandas and NumPy libraries to manipulate the data set.

NumPy

#Using the sections below, we can check the library version to determine we are not using an old version

```
import numpy as np

np.__version__

'1.12.1'

#The following statements create a list with numbers from 0 to 9

L = list(range(10))

#The next few lines of code convert integers to strings. This process is called list comprehension. This process is one of the best ways to handle list manipulations.

[str(c) for c in L]

['0', '1', '2', '3', '4', '5', '6', '7', '8', '9']
```

OceanofPDF.com

[type(item) for item in L]

[int, int, int, int, int, int, int, int, int, int]

Create Arrays

Arrays are homogeneous types of data. If you are familiar with programming languages, you will have an idea of how you can use arrays. Arrays only hold specific variables in the data set, and this is true in any programming language.

#creating arrays

np.zeros(10, dtype='int')

array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0])

#We will now create a 3 x 5 array

np.ones((3,5), dtype=float)

array([[1., 1., 1., 1., 1.],

1., 1., 1., 1., 1.],

1., 1., 1., 1., 1.]])

#The next lines of code are used to create arrays with a predefined value np.full((3,5),1.23)

array([[1.23, 1.23, 1.23, 1.23, 1.23],

1.23, 1.23, 1.23, 1.23, 1.23],

1.23, 1.23, 1.23, 1.23, 1.23]]) #We are creating an array using the sequence np.arange(0, 20, 2)

OceanofPDF.com

array([0, 2, 4, 6, 8,10,12,14,16,18])

#The next few lines of code are used to create arrays with variables with the space between the values np.linspace(0, 1, 5)

array([0., 0.25, 0.5 , 0.75, 1.])

#We will now create a 3 x 3 array with the values of mean and standard deviation as 0 and 1, respectively. We will use the dimension np.random.normal(0, 1, (3,3))

array([[0.72432142, -0.90024075, 0.27363808],

0.88426129, 1.45096856, -1.03547109], [-0.42930994, -1.02284441, -1.59753603]]) #create an identity matrix

np.eye(3) array([[1., 0., 0.],

0., 1., 0.],

0., 0., 1.]])

#set a random seed np.random.seed(0)

x1 = np.random.randint(10, size=6) #one dimension

x2 = np.random.randint(10, size=(3,4)) #two dimension

x3 = np.random.randint(10, size=(3,4,5)) #three dimension

print("x3 ndim:", x3.ndim)

print("x3 shape:", x3.shape)

print("x3 size: ", x3.size)

('x3 ndim:', 3)

OceanofPDF.com

('x3 shape:', (3, 4, 5))

('x3 size: ', 60)

Array Indexing

It is important to remember the index of the array only begins at zero. The first value in an array has the index zero.

x1 = np.array([4, 3, 4, 4, 8, 4])

x1

array([4, 3, 4, 4, 8, 4])

#assess value to index zero

x1[0]

4

#assess fifth value

x1[4]

8

#get the last value

x1[-1]

4

#get the second last value

OceanofPDF.com

```
x1[-2]
8
#We are specifying the rows and columns for a multidimensional array
array([[3, 7, 5, 5],
[0, 1, 5, 9],
[3, 0, 5, 0]])
#1st row and 2nd column value
x2[2,3]
0
#3rd row and last value from the 3rd column
x2[2,-1]
0
#replace value at 0,0 index
x2[0,0] = 12
x2
array([[12, 7, 5, 5],
0, 1, 5, 9],
3, 0, 5, 0]])
```

OceanofPDF.com

Array Slicing

Most programming languages allow you to access a range of elements in the array. To do this, you would need to slice the array into parts.

```
x = np.arange(10)

x

array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])

#from the first to the fourth position

x[:5]

array([0, 1, 2, 3, 4])

#from the fourth to the last element

x[4:]

array([4, 5, 6, 7, 8, 9])

#from the fourth to the sixth position

x[4:7]

array([4, 5, 6])

#The code returns the elements at even indices

x[ : : 2]

array([0, 2, 4, 6, 8])

#These lines return elements in intervals of two from the first element
```

OceanofPDF.com

x[1::2]

array([1, 3, 5, 7, 9])

#reverse the array

x[::-1]

array([9, 8, 7, 6, 5, 4, 3, 2, 1, 0])

Array Concatenation

Most programmers split the data set into multiple arrays, but they would need to combine or concatenate these arrays to perform different operations. You do not have to type the elements in the different arrays but can combine them to handle the operations easily.

#Using the lines of code below, you can concatenate more than two arrays at once

x = np.array([1, 2, 3])

y = np.array([3, 2, 1])

z = [21,21,21]

np.concatenate([x, y, z])

array([1, 2, 3, 3, 2, 1, 21, 21, 21])

#The following functions can be used to create two-dimensional arrays

grid = np.array([[1,2,3],[4,5,6]])

np.concatenate([grid,grid])

OceanofPDF.com

array([[1, 2, 3],

[4, 5, 6],

[1, 2, 3],

[4, 5, 6]])

#You can define the column and row wise matrix using the axis parameter np.concatenate([grid,grid],axis=1)

array([[1, 2, 3, 1, 2, 3],

[4, 5, 6, 4, 5, 6]])

In the code written above, the concatenation function has been used on the arrays. It is important to note that the arrays with the same data types are combined for these operations.

You can create different types of arrays based on the number of dimensions. But how do you combine arrays of different dimensions together? The next few lines of code use the np.vstack and the np.hstack (similar to the VLookUp and HLookUp functions in excel) to combine the data.

x = np.array([3,4,5])

grid = np.array([[1,2,3],[17,18,19]])

np.vstack([x,grid])

array([[3, 4, 5],

1, 2, 3], [17, 18, 19]])

#The following keywords and functions can be used to add arrays np.hstack
z = np.array([[9],[9]])

OceanofPDF.com

np.hstack([grid,z]) array([[1, 2, 3, 9], [17, 18, 19, 9]])

You can split the array using pre-defined positions or criteria.

x = np.arange(10)

x

array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9]) x1,x2,x3 = np.split(x,[3,6]) print x1,x2,x3

[0 1 2] [3 4 5] [6 7 8 9]

grid = np.arange(16).reshape((4,4)) grid

upper,lower = np.vsplit(grid,[2]) print (upper, lower)

(array([[0, 1, 2, 3],

[4, 5, 6, 7]]), array([[8, 9, 10, 11], [12, 13, 14, 15]]))

The NumPy directory in Python gives you access to different mathematical functions, which can be used in Python. Some of these functions include, sum, divide, abs, multiple, mod, power, log, sin, tan, cos, mean, min, max, var, etc.

You can use one or more of these functions to perform arithmetic calculations. If you want to learn more about the use of these functions, you can read the NumPy documentation to learn more about the various functions.

Now that we have looked at how we can use the NumPy library let us look at the Pandas library and see how it can be used. Make sure to read each line of code before you manipulate the information you have collected.

OceanofPDF.com

Let's Start with Pandas

import pandas as pd

#We will create a data frame dictionary to cover the names of rows and columns

data = pd.DataFrame({'Country': ['Russia','Colombia','Chile','Ecuador','Nigeria'],

Rank':[121,40,100,130,11]})

data

CountryRank

0Russia121

1Colombia40

2Chile100

3Equador130

4Nigeria11

#The following lines of code produce an analysis of the data

data.describe()

Rank

count5.000000

mean80.400000

std52.300096

OceanofPDF.com

min11.000000

25%40.000000

50%100.000000

75%121.000000

max130.000000

How to Summarize Data Using Pandas

The Pandas library has different functions that enable you to obtain a statistical summary of all the data set variables. You can do this using the describe function or method in the library. If you want to look at all the information present in the data set, you can use the function or method info().

#Among other things, it shows the data set has five rows and two columns with their respective names. data.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 5 entries, 0 to 4

Data columns (total two columns):

Country 5 non-null object

Rank 5 non-null int64

dtypes: int64(1), object(1)

memory usage: 152.0+ bytes

#We will now create another data frame

OceanofPDF.com

data = pd.DataFrame({'group':['a', 'a', 'a', 'b', 'b', 'b', 'c', 'c', 'c'],'ounces':[4, 3, 12, 6, 7.5, 8, 3, 5, 6]})

data

groupounces

0a4.0

1a3.0

2a12.0

3b6.0

4b7.5

5b8.0

6c3.0

7c5.0

8c6.0

#We will sort the data frame by making changes to the data set using the following functions: inplace = True and data.sort_values(by= ['ounces'],ascending=True,inplace=False) groupounces

1a3.0

6c3.0

0a4.0

7c5.0

3b6.0

OceanofPDF.com

8c6.0

4b7.5

5b8.0

2a12.0

You can now sort the data using different columns.

data.sort_values(by=['group','ounces'],ascending=
[True,False],inplace=False)

groupounces

2a12.0

0a4.0

1a3.0

5b8.0

4b7.5

3b6.0

8c6.0

7c5.0

6c3.0

How to Work with Noise or Duplicate Records

Every data set has duplicate columns and rows that are often termed noise. As mentioned earlier, you need to get rid of this noise if you want to obtain accurate results from the analysis. Never feed a model with unprocessed data. This section sheds some light on how you can remove any duplicates in the data set.

#create another data with duplicated rows

```
data = pd.DataFrame({'k1':['one']*3 + ['two']*4, 'k2':[3,2,1,3,3,4,4]})
data
k1k2
0one3
1one2
2one1
3two3
4two3
5two4
6two4
#sort values
data.sort_values(by='k2')
k1k2
```

OceanofPDF.com

2one1

1one2

0one3

3two3

4two3

5two4

6two4

#remove duplicates - ta-da!

data.drop_duplicates()

k1k2

0one3

1one2

2one1

3two3

5two4

In the above example, we are mapping the values present in the columns and rows to remove any duplicate values present in the data set. You can remove these duplicates using certain columns as parameters. The following lines of code look at how we can remove duplicates from specific columns.

data.drop_duplicates(subset='k1')

k1k2

OceanofPDF.com

0one3

3two3

Categorization of Data

Now that you have removed the duplicates and noise in the data set let us look at how you can categorize the data based on some predefined criteria or rules. You need to do this often if you want to process the data into categories before you run the data through the model. Let us assume you work on data that has information for each country. You may want to split the data based on the continents. If you want to do this, you should create a new variable. The code below can be used to achieve this.

data = pd.DataFrame({'food': ['bacon', 'pulled pork', 'bacon', 'Pastrami','corned beef', 'Bacon', 'pastrami', 'honey ham', 'nova lox'],

'ounces': [4, 3, 12, 6, 7.5, 8, 3, 5, 6]})

data

foodounces

0bacon4.0

1pulled pork3.0

2bacon12.0

3Pastrami6.0

4corned beef7.5

5Bacon8.0

OceanofPDF.com

6pastrami3.0

7honey ham5.0

8nova lox6.0

Let us now try to map one animal to another based on each animal's source of food. To do this, we will create a new variable. This variable will store the mapping. If you want the model to perform the mapping, you need to create a dictionary to map the food to the animals. Use a map function to pull every variable from the dictionary. The following lines of code will help you do this.

meat_to_animal = {

'bacon': 'pig',

'pulled pork': 'pig',

'pastrami': 'cow',

'corned beef': 'cow',

'honey ham': 'pig',

'nova lox': 'salmon'

}

def meat_2_animal(series):

if series['food'] == 'bacon':

return 'pig'

elif series['food'] == 'pulled pork':

OceanofPDF.com

```
return 'pig'
elif series['food'] == 'pastrami':
return 'cow'
elif series['food'] == 'corned beef':
return 'cow'
elif series['food'] == 'honey ham':
return 'pig'
else:
return 'salmon'
#create a new variable
data['animal']= data['food'].map(str.lower).map(meat_to_animal)
data
```

	food	ounces	animal
0	bacon	4.0	pig
1	pulled pork	3.0	pig
2	bacon	12.0	pig
3	Pastrami	6.0	cow
4	corned beef	7.5	cow
5	Bacon	8.0	pig

OceanofPDF.com

6pastrami3.0cow

7honey ham5.0pig

8nova lox6.0salmon

#Alternatively, you can apply the function once you convert the food values into lower case

lower = lambda x: x.lower()

data['food'] = data['food'].apply(lower)

data['animal2'] = data.apply(meat_2_animal, axis='columns')

data

foodouncesanimalanimal2

0bacon4.0pigpig

1pulled pork3.0pigpig

2bacon12.0pigpig

3pastrami6.0cowcow

4corned beef7.5cowcow

5bacon8.0pigpig

6pastrami3.0cowcow

7honey ham5.0pigpig

8nova lox6.0salmonsalmon

OceanofPDF.com

You can also use the assign function to create new variables. This function uses different Pandas functions that you can use while you work with different data sets. You need to keep these functions in mind.

How to Solve Problems Using Pandas

The next section sheds some light on how you can use the Pandas library to solve different problems. The Pandas library has different packages that allow you to handle big data sets.

data.assign(new_variable = data['ounces']*10)

foodouncesanimalanimal2new_variable

0bacon4.0pigpig40.0

1pulled pork3.0pigpig30.0

2bacon12.0pigpig120.0

3pastrami6.0cowcow60.0

4corned beef7.5cowcow75.0

5bacon8.0pigpig80.0

6pastrami3.0cowcow30.0

7honey ham5.0pigpig50.0

8nova lox6.0salmonsalmon60.0

Let us now remove the animal2 column from the data frame.

OceanofPDF.com

```
data.drop('animal2',axis='columns',inplace=True)
```

```
data
```

foodouncesanimal

0bacon4.0pig

1pulled pork3.0pig

2bacon12.0pig

3Pastrami6.0cow

4corned beef7.5cow

5Bacon8.0pig

6pastrami3.0cow

7honey ham5.0pig

8nova lox6.0salmon

How to Clean Data Sets

Most organizations use different sources to collect information. These data sets will need to be cleaned and processed. You may find some missing values in the data set. If you want to remove any missing data from the data set, you need to substitute the missing value with a dummy variable.

You can also use a default value if you would like to. There is a possibility there are outliers in the data set, and you need to remove them. Let us see how this can be done in the directory.

OceanofPDF.com

#In the sections below, we will use a series function from the library to create an array

data = pd.Series([1., -999., 2., -999., -1000., 3.])

data

0 1.0

-999.0

2 2.0

3 -999.0

4 -1000.0

5 3.0

dtype: float64

#replace -999 with NaN values data.replace(-999, np.nan,inplace=True) data

0 1.0

NaN

2.0

3 NaN

4 -1000.0

5 3.0

dtype: float64

OceanofPDF.com

#The following lines of code allow you to replace multiple values at once. data = pd.Series([1., -999., 2., -999., -1000., 3.]) data.replace([-999,-1000],np.nan,inplace=True) data

0 1.0

1 NaN

2.0

NaN

4 NaN

5 3.0

dtype: float64

The next few lines of code will help you rename columns and rows.

data = pd.DataFrame(np.arange(12).reshape((3, 4)),index=['Ohio', 'Colorado', 'New York'],columns=['one', 'two', 'three', 'four'])

data

onetwothreefour

Ohio0123

Colorado4567

New York891011

Using rename function

data.rename(index = {'Ohio':'SanF'},

OceanofPDF.com

columns={'one':'one_p','two':'two_p'},inplace=True) data

one_ptwo_pthreefour

SanF0123

Colorado4567

New York891011

#We can perform the same functions using string

data.rename(index = str.upper, columns=str.title,inplace=True)

data

One_pTwo_pThreeFour

SANF0123

COLORADO4567

NEW YORK891011

Now, we need to categorize or split the continuous variables.

ages = [20, 22, 25, 27, 21, 23, 37, 31, 61, 45, 41, 32]

we will now divide the large data set into smaller bins or segments.

bins = [18, 25, 35, 60, 100]

cats = pd.cut(ages, bins)

cats

OceanofPDF.com

[(18, 25], (18, 25], (18, 25], (25, 35], (18, 25], ..., (25, 35], (60, 100], (35, 60], (35, 60], (25, 35]] Length: 12

Categories (4, object): [(18, 25] < (25, 35] < (35, 60] < (60, 100]] #This statement includes the bin values to the right pd.cut(ages,bins,right=False)

[[18, 25), [18, 25), [25, 35), [25, 35), [18, 25), ..., [25, 35), [60, 100), [35, 60), [35, 60), [25, 35)] Length: 12

Categories (4, object): [[18, 25) < [25, 35) < [35, 60) < [60, 100)] #The pandas library assigns each categorical variable with an encoder. cats.labels

array([0, 0, 0, 1, 0, 0, 2, 1, 3, 2, 2, 1], dtype=int8)

#The following will show the number of observations falling under a certain range

pd.value_counts(cats)

(18, 25] 5

(35, 60] 3

(25, 35] 3

(60, 100] 1

dtype: int64

We can also pass a name to every label.

bin_names = ['Youth', 'YoungAdult', 'MiddleAge', 'Senior'] new_cats = pd.cut(ages, bins,labels=bin_names)

pd.value_counts(new_cats)

OceanofPDF.com

Youth5

MiddleAge3

YoungAdult3

Senior1

dtype: int64

#using the following lines of code, you can calculate the cumulative sum

pd.value_counts(new_cats).cumsum()

Youth5

MiddleAge3

YoungAdult3

Senior1

dtype: int64

The pandas directory also has libraries which allow you to create pivots or group variables into clusters. It offers easy-to-use pivot tables to perform these forms of clustering. It is for this reason you need to understand and learn how you can do this.

df = pd.DataFrame({'key1' : ['a', 'a', 'b', 'b', 'a'],

'key2' : ['one', 'two', 'one', 'two', 'one'],

'data1' : np.random.randn(5),

'data2' : np.random.randn(5)})

OceanofPDF.com

df

data1data2key1key2

00.9735990.001761a

10.207283-0.990160a

21.0996421.872394b

30.939897-0.241074b

40.6063890.053345a

#calculate the mean of data1 column by key1

grouped = df['data1'].groupby(df['key1'])

grouped.mean()

key1

0.595757 b 1.019769

Name: data1, dtype: float64

#We will now slice the data frame

dates = pd.date_range('20130101',periods=6)

df = pd.DataFrame(np.random.randn(6,4),index=dates,columns=list('ABCD')) df

ABCD

2013-01-011.030816-1.2769890.837720-1.490111 2013-01-02-1.070215-0.2091290.604572-1.743058

OceanofPDF.com

2013-01-031.5242271.8635751.2913781.300696

2013-01-040.918203-0.158800-0.964063-1.990779

2013-01-050.0897310.114854-0.5858150.298772

2013-01-060.2222600.435183-0.0457480.049898 #this line gets the first n rows from the data set df[:3]

ABCD

2013-01-011.030816-1.2769890.837720-1.490111

2013-01-02-1.070215-0.2091290.604572-1.743058

2013-01-031.5242271.8635751.2913781.300696

#the array is sliced based on the dates

df['20130101':'20130104']

ABCD

2013-01-011.030816-1.2769890.837720-1.490111

2013-01-02-1.070215-0.2091290.604572-1.743058

2013-01-031.5242271.8635751.2913781.300696

2013-01-040.918203-0.158800-0.964063-1.990779 #This line is used to slice the code based on column names df.loc[:,['A','B']]

AB

2013-01-011.030816-1.276989

2013-01-02-1.070215-0.209129

OceanofPDF.com

2013-01-031.5242271.863575

2013-01-040.918203-0.158800

2013-01-050.0897310.114854

2013-01-060.2222600.435183

#we now slice the array based on row and column labels
df.loc['20130102':'20130103',['A','B']] AB

2013-01-02-1.070215-0.209129

2013-01-031.5242271.863575

#we are slicing the array based on column indices

df.iloc[3] #returns 4th row (index is 3rd)

0.918203 B -0.158800 C -0.964063 D -1.990779

Name: 2013-01-04 00:00:00, dtype: float64 #this line of code returns a specific range df.iloc[2:4, 0:2]

AB

2013-01-031.5242271.863575

2013-01-040.918203-0.158800

#the following lines of code return specific columns and rows using lists with row and columns indices

df.iloc[[1,5],[0,2]]

AC

2013-01-02-1.0702150.604572

OceanofPDF.com

2013-01-060.222260-0.045748

Using Pre-Defined Conditions

You can also perform Boolean indexing using different values present in the columns. This is the only way you can filter the data set based on some predefined conditions.

df[df.A > 1]

ABCD

2013-01-011.030816-1.2769890.837720-1.490111

2013-01-031.5242271.8635751.2913781.300696

#we will now copy the data

df2 = df.copy()

df2['E']=['one', 'one','two','three','four','three']

df2

ABCDE

2013-01-011.030816-1.2769890.837720-1.490111one

2013-01-02-1.070215-0.2091290.604572-1.743058one

2013-01-031.5242271.8635751.2913781.300696two

2013-01-040.918203-0.158800-0.964063-1.990779three

2013-01-050.0897310.114854-0.5858150.298772four

2013-01-060.2222600.435183-0.0457480.049898three #this line selects rows based on the values in columns

OceanofPDF.com

df2[df2['E'].isin(['two','four'])] ABCDE

2013-01-031.5242271.8635751.2913781.300696two

2013-01-050.0897310.114854-0.5858150.298772four #this line selects those rows without two and four df2[~df2['E'].isin(['two','four'])] ABCDE

2013-01-011.030816-1.2769890.837720-1.490111one

2013-01-02-1.070215-0.2091290.604572-1.743058one

2013-01-040.918203-0.158800-0.964063-1.990779three

2013-01-060.2222600.435183-0.0457480.049898three

The query method in the pandas library allows you to select the required columns in the data set. You only need one criterion to perform this function.

#This is a list of columns where the value A is greater than C

df.query('A > C')

ABCD

2013-01-011.030816-1.2769890.837720-1.490111

2013-01-031.5242271.8635751.2913781.300696

2013-01-040.918203-0.158800-0.964063-1.990779

2013-01-050.0897310.114854-0.5858150.298772

2013-01-060.2222600.435183-0.0457480.049898

#using OR condition

OceanofPDF.com

df.query('A < B | C > A')

ABCD

2013-01-02-1.070215-0.2091290.604572-1.743058

2013-01-031.5242271.8635751.2913781.300696

2013-01-050.0897310.114854-0.5858150.298772

2013-01-060.2222600.435183-0.0457480.049898

Pivot tables can be used to customize the information in an easily readable manner. You can use pivot tables to learn more about the data set. Some people use Excel to build and develop pivot tables to better analyze and understand the data.

#create a data frame

data = pd.DataFrame({'group': ['a', 'a', 'a', 'b','b', 'b', 'c', 'c','c'],

'ounces': [4, 3, 12, 6, 7.5, 8, 3, 5, 6]})

data

groupounces

0a4.0

1a3.0

2a12.0

3b6.0

4b7.5

5b8.0

OceanofPDF.com

6c3.0

7c5.0

8c6.0

The following statements calculate he mean of the groups

data.pivot_table(values='ounces',index='group',aggfunc=np.mean)

group

6.333333 b 7.166667 c 4.666667

Name: ounces, dtype: float64 #calculate count by each group

data.pivot_table(values='ounces',index='group',aggfunc='count') group

3

3 c 3

Name: ounces, dtype: int64

Let us now look at a real-world example and see how the Pandas and NumPy libraries can be used.

OceanofPDF.com

Chapter Seven: Exploring the Data Set

In this chapter, we will use the different functions listed above to obtain information about adults. To work with the same data set, you can download it from the following website: **https://s3-ap-southeast-1.amazonaws.com/he-public-data/datafiles19cdaf8.zip**. This data has data that is classified using binary conditions. This example aims to calculate the salary of a person using specific variables in the data set.

#load the data

train = pd.read_csv("~/Adult/train.csv")

test = pd.read_csv("~/Adult/test.csv")

#check data set

train.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 32561 entries, 0 to 32560

Data columns (total 15 columns):

age 32561 non-null int64

workclass 30725 non-null object

fnlwgt 32561 non-null int64

OceanofPDF.com

education 32561 non-null object

education.num 32561 non-null int64

marital.status 32561 non-null object

occupation 30718 non-null object

relationship 32561 non-null object

race 32561 non-null object

sex 32561 non-null object

capital.gain 32561 non-null int64

capital.loss 32561 non-null int64

hours.per.week 32561 non-null int64

native.country 31978 non-null object

target 32561 non-null object

dtypes: int64(6), object(9)

memory usage: 3.7+ MB

We are going to split the data into two parts – the training and testing data sets. The training data set we will use has 32561 rows and 15 columns. Of these fifteen columns, only six of them have integer values, while the other have character or object data. You can check the values and records present in the testing data set in a similar manner. The number of columns and rows in the data set can be defined using the lines of code written below:

```
print ("The train data has",train.shape)
```

OceanofPDF.com

print ("The test data has",test.shape)

('The train data has', (32561, 15))

('The test data has', (16281, 15))

#Let have a glimpse of the data set

train.head()

ageworkclassfnlwgteducationeducation.nummarital.statusoccupationrelationshipracesexcapital.gaincapit

al.losshours.per.weeknative.countrytarget

039State-gov77516Bachelors13Never-marriedAdm-clericalNot-in-familyWhiteMale2174040United-States<=50K

150Self-emp-not-inc83311Bachelors13Married-civ-spouseExec-managerialHusbandWhiteMale0013United-States<=50K

238Private215646HS-grad9DivorcedHandlers-cleanersNot-in-familyWhiteMale0040United-States<=50K

353Private23472111th7Married-civ-spouseHandlers-cleanersHusbandBlackMale0040United-States<=50K

428Private338409Bachelors13Married-civ-spouseProf-specialtyWifeBlackFemale0040Cuba<=50K Let us verify if the data set has any missing values. nans = train.shape[0] - train.dropna().shape[0]

print ("%d rows have missing values in the train data" %nans)

nand = test.shape[0] - test.dropna().shape[0]

OceanofPDF.com

print ("%d rows have missing values in the test data" %nand)

When you run the above lines of code, you will see there are 2399 rows in the training data set with missing values. There are 1221 rows in the test data set with missing values. Let us now look for the columns with missing values.

#only 3 columns have missing values

train.isnull().sum()

age0

workclass1836

fnlwgt0

education0

education.num0

marital.status0

occupation1843

relationship0

race0

sex0

capital.gain0

capital.loss0

hours.per.week0

native.country583

OceanofPDF.com

target0

dtype: int64

When you use these character variables, you can count the different values in the data set. These values will be unique.

cat = train.select_dtypes(include=['O'])

cat.apply(pd.Series.nunique)

workclass8

education16

marital.status7

occupation14

relationship6

race5

sex2

native.country41

target2

dtype: int64

From the previous section, we learned how you can work with missing values in the data set. Let us now see how you can substitute the missing values from different values.

#Education

train.workclass.value_counts(sort=True)

OceanofPDF.com

```
train.workclass.fillna('Private',inplace=True)
#Occupation
train.occupation.value_counts(sort=True)
train.occupation.fillna('Prof-specialty',inplace=True)
#Native Country
train['native.country'].value_counts(sort=True)
train['native.country'].fillna('United-States',inplace=True)
train.isnull().sum()
age0
workclass0
fnlwgt0
education0
education.num0
marital.status0
occupation0
relationship0
race0
sex0
capital.gain0
capital.loss0
```

OceanofPDF.com

hours.per.week0

native.country0

target0

dtype: int64

We will look at the target variables to see if there are any issues in the data set.

#check proportion of target variable

train.target.value_counts()/train.shape[0]

<=50K 0.75919

>50K 0.24081

Name: target, dtype: float64

pd.crosstab(train.education, train.target,margins=True)/train.shape[0]

target<=50K>50KAll

education

10th0.0267500.0019040.028654

11th0.0342430.0018430.036086

12th0.0122850.0010130.013298

1st-4th0.0049750.0001840.005160

5th-6th0.0097360.0004910.010227

7th-8th0.0186110.0012280.019840

OceanofPDF.com

9th0.0149570.0008290.015786

Assoc-acdm0.0246310.0081390.032769

Assoc-voc0.0313570.0110870.042443

Bachelors0.0962500.0682100.164461

Doctorate0.0032860.0093980.012684

HS-grad0.2710600.0514420.322502

Masters0.0234640.0294520.052916

Preschool0.0015660.0000000.001566

Prof-school0.0046990.0129910.017690

Some-college0.1813210.0425970.223918

All0.7591900.2408101.000000

From the data above, you will see that 75% of the people in the data set have a salary of above $50,000 and 27% of these individuals have passed high school. This is accurate since people with lower levels of education will not make a salary over $50,000.

#The following statements are to load sklearn to encode the variables using an object type

from sklearn import preprocessing

for x in train.columns:

if train[x].dtype == 'object':

lbl = preprocessing.LabelEncoder()

OceanofPDF.com

lbl.fit(list(train[x].values))

train[x] = lbl.transform(list(train[x].values))

We will now look at the different changes that have been applied to the data set.

train.head()

ageworkclassfnlwgteducationeducation.nummarital.statusoccupationrelationshipracesexcapital.gaincapit al.losshours.per.weeknative.countrytarget

039677516913401412174040380

150583311913230410013380

238321564611905141004038 0

353323472117250210040380

428333840991329520004040

If you pay attention to the output of the above code, you will see that the variables have been converted into numeric data types.

#<50K = 0 and >50K = 1

train.target.value_counts()

24720

1 7841

Name: target, dtype: int64 Building a Random Forest Model

OceanofPDF.com

We are now going to test the accuracy of the model using a random forest model.

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.cross_validation import cross_val_score

from sklearn.metrics import accuracy_score

y = train['target'] del train['target'] X = train

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3,random_state=1,stratify=y) #train the RF classifier

clf = RandomForestClassifier(n_estimators = 500, max_depth = 6) clf.fit(X_train,y_train)

RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini', max_depth=6, max_features='auto', max_leaf_nodes=None, min_impurity_split=1e-07, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=500, n_jobs=1, oob_score=False, random_state=None, verbose=0, warm_start=False)

clf.predict(X_test)

We will now check the accuracy of the model on the basis of some predictions that are want to make about the data set.

#make prediction and check model's accuracy

prediction = clf.predict(X_test)

OceanofPDF.com

```
acc = accuracy_score(np.array(y_test),prediction)

print ('The accuracy of Random Forest is {}'.format(acc))
```

The accuracy of Random Forest is 0.85198075545.

The algorithm we used above will give you results with an 85% accuracy. If you want to improve the model's accuracy, you need to tweak the program and functions used. We are dividing the data set into two parts – the testing and training data set. We use the training data set to enable the model to predict the required outcomes. We then run the testing data set through the model to see if it performs well and predicts an output similar to the training data set.

OceanofPDF.com

Chapter Eight: How to Summarize Data with Python

For the purpose of this chapter, we will use the data set for supermarket sales. You can download this data using the following link:

https://www.kaggle.com/aungpyaeap/supermarket-sales/data.

Download this information and save it on your local drive for use. You can import the information onto python using the following lines of code:

```
#import library

import pandas as pd

#import file

ss = pd.read_csv('supermarket_sales.csv')

#preview data

ss.head()
```

Obtain Information About the Data Set

You can use the following functions to obtain information about the data set.

OceanofPDF.com

Info()

This function gives you a concise summary of the data frame extracted from the data set you have uploaded. This function can be used if you use the Pandas library. This function, among other functions, can be used to perform data cleaning operations. The function provides all the information you need about the data set, such as record counts, columns, data types, memory usage, names of columns and rows, and the index range. When you look at the summary provided for the above data set, you can obtain the following information:

- The record count is 1000
- There are 17 columns in the data set
- You can update the column names if you want to remove any white paces
- Identify the data types
- The date and time columns in the data set use the object type dtype. You need to update these to match the required formats
- The data set does not have any missing values

Describe()

This function gives you the descriptive statistics which will provide you the information you need about the distribution the data set follows.

This function does not look at the NaN values, and it can be used on a specific series of data or data frames.

OceanofPDF.com

The results will only be for the numeric data types, but you can use the function include if you want to obtain some statistics for every non-numeric data type in the data set. If you only want to see the results for specific data frames, you can use the exclude function. The function, by default, will return the percentile values. The percentiles extracted are in the 25^{th}, $50^{th,}$ and 75^{th} range. If you want to update the different percentiles obtained, you can add the percentiles using the percentiles function. The values used in this function should lie between the values 0 and 1. This function will apply to both series and data frames.

Value_counts()

This function returns unique values for the series specified in the function. This function also excludes NaN values. This function's results are in descending order, but you can always change the way the data is presented using the function **ascending = true**. If you want to obtain the frequency of every value in the data set, you need to set the function **normalize = true**.

nunique()

You can use this function if you count the number of distinct observations. You can use it on series or data frames. This function, by default, excludes the NaN values. If you want to use the NaN values in the data set, you need to use the parameter **dropna = false**. Since there are no missing values in the data set, let us create a data frame to understand how this function works on data sets with NaN values.

d = {'A': [1, 2, None], 'B': [3, 4, 2], 'C': [3, None, None]}

OceanofPDF.com

```
df = pd.DataFrame(d)

df
```

When you run the dropna function, you can extract the parameters without the NaN values.

Sun()

This function is used to obtain the sum of all the values of the axis requested in the function. You can use this function on both series and data frames. If you want to apply this function only to numeric data types, you can use the parameter **numeric_only = True**. You can also use the min_count parameter if you want to set a condition to let the compiler know when you can apply the function. We will create another data frame, which you can use to demonstrate how this parameter works.

```
d = {'A': [1, 2, None, 5, 8], 'B': [3, 4, 2, 4, 5],

    'C': [3, None, None, 3, 4]}

df = pd.DataFrame(d)

df
```

count()

This function returns the number of null or non-null observations, and you can apply this function on both series and data frames. If you want to obtain the information only on numeric data types, you should use the parameter **numeric_only = True**.

OceanofPDF.com

Minimum, Maximum, Mean, and Median

The following are the functions you can use to extract the minimum, maximum, mean, and median values.

- Min()
- Max()
- Mean()
- Median()

You can apply these functions to both series and data frames.

In this chapter, we looked at some functions that help you summarize the data using various library functions. The info() function will allow you to obtain information about the data frame's format and structure. You can also use the describe() function to obtain descriptive statistics information about the data set. We then looked at how you can obtain the number of records per column and other functions you can use.

OceanofPDF.com

Chapter Nine: Steps to Build Data Analysis Models in Python

You have come quite far now and have learned the basics of data mining and data analysis. We will now look at how you can use Python to develop a predictive analytics model. We will see how you can divide the data into a training and testing data set and use it to develop and measure the model's performance. When you build any model, you need to spend enough time understanding the business need and developing the hypothesis based on that need. To do this, you need to understand the domain. When you do this, you can identify the problems and queries and map the data to them. This is the only way you can design the business solution to the problem or query. The following are some reasons why you need to do this:

- You can spend enough time to learn more about the data analysis process
- You will not have any other perspective or thoughts which could lead to a bias

This ensures that the hypothesis you develop will not have any bias.

You need to spend sufficient time to complete this stage, and therefore, you need to make this your standard practice. If you perform this step carefully, you will not have to work on multiple iterations over the course of the development of the model. Let us look at the different steps followed in this process. We will look at the time you spend during each stage.

OceanofPDF.com

- The first thing you need to do is perform a descriptive analysis of the data you have collected. This will take a majority of your time
- The second step includes the cleaning and treating of data. In this step, you look for any missing values and outliers in the data set. You may spend close to 40% of your time doing this, so make sure you obtain the right data set
- When you obtain the required data set, you need to start the data modeling process. Since you have cleaned and processed the data, you can use different algorithms to develop the model
- The last step is to estimate the model and its performance

This is the time you spend when you develop the model for the first time. You may spend a lot of time to work through the model and make the necessary updates. We will go through each step to see what the step entails and how much time you should spend on it.

Step One: Descriptive Analysis and Data Exploration

If you are starting off as a data scientist or analyst, this may take you some time. When you get the hang of the process, you can use the different techniques to automate numerous operations. It will take you a long time to prepare the data you collected before analyzing or interpreting the results. It is because of this people use different libraries in Python to automate these processes.

OceanofPDF.com

We looked at some of these processes in the previous chapters in the book. Some people use advanced deep learning and machine learning tools to improve the task they are performing. These tools also reduce the time you take to perform the task. Do not use complex concepts, such as feature engineering, if you want to reduce the time you spend performing this task. The objective of this step is to identify any missing values or features in the data set. You can perform the following steps to achieve this objective:

- Identify the ID, target features, and inputs in the data set
- Differentiate between categorical and numerical features
- Identify the columns with missing values

Step Two: Remove Missing Values in the Data Set

There are numerous ways to work with missing values in a data set. We looked at these in great detail in the previous chapters. The following are some smart and quick techniques, and these will help you develop effective models.

- You can create a dummy value or flag to replace any missing values in the data set. It is best to use this technique only when there is minimal or no information about the missing values in the data set
- You can use measures of central tendency, such as mean, median, and mode, to replace any missing values in the data set. The mean and median measures perform extremely well. Experts use the mean to replace any missing values regardless of the data they are dealing with.

OceanofPDF.com

Having said that, if the data follows a skewed distribution, you should use the median.

- It is best to replace missing values in the data set using different variables in the data set. If you know the type of data you are using, you can create new labels that can be used to replace any missing values in the data set using categorical variables. You can also use a frequency mix to replace any missing values in the data set using a value with higher frequencies.

It is best to reduce the time you spend on this activity by five minutes. Therefore, you need to use simple methods.

Step Three: Data Modeling

You can use different techniques, such as GBM or random forest modeling, to develop the data model. You need to select the model you want to use depending on the problem or query you want to address. These models are the best way to develop or create benchmark solutions. Most data scientists or analysts use any of the above methods to develop their models. You can continue to use the same techniques to develop the model.

Step Four: Estimate the Model's Performance

Once you develop a model and write the code to build it, you need to use different methods to estimate the performance of the model. You should first divide the data set into a training and testing data set. Using the testing data set, you can determine the accuracy of the model. It is best to split the data into the training and testing data set using the 70:30 ratio. This is extremely easy to do, and we will look at how you can do this in the next chapter.

OceanofPDF.com

Chapter Ten: How to Build the Model

We will put the techniques we have discussed above into action in this chapter. We will build the required model using Python.

Step One: Import Libraries Required to Develop the Model and Perform the Analysis. Read the Information and Split the Data into Training and Testing Data Sets.

```
import pandas as pd

import numpy as np

from sklearn.preprocessing import LabelEncoder

import random

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import GradientBoostingClassifier

train=pd.read_csv('C:/Users/Analytics
Vidhya/Desktop/challenge/Train.csv')

test=pd.read_csv('C:/Users/Analytics Vidhya/Desktop/challenge/Test.csv')

train['Type']='Train' #These lines create a flag for the test and train data sets

test['Type']='Test'
```

OceanofPDF.com

fullData = pd.concat([train,test],axis=0) #We are combining the training and testing data sets

Step Two: Summarize the Data Set into Columns

fullData.columns #This gives you the column headers as the output

fullData.head(10) #This shows the first ten records in your frame

fullData.describe() #This shows you the statistical summary of the data set

Step Three: Identification of Variables

In this step, we will identify the following:

- ID Variables
- Target Variables
- Categorical Variables
- Numerical Variables
- Other Variables

ID_col = ['REF_NO']

target_col = ["Account.Status"]

cat_cols = ['children','age_band','status','occupation','occupation_partner','home_status', 'family_income','self_employed', 'self_employed_partner','year_last_moved','TVarea','post_code','post_area','gender','region']

OceanofPDF.com

num_cols= list(set(list(fullData.columns))-set(cat_cols)-set(ID_col)-set(target_col)-set(data_col))

other_col=['Type'] #Test and Train Data set identifier

Step Four: Identify the Missing Values in the Data Set and Create a Flag

fullData.isnull().any()#This function returns the values true and false. The former indicates a missing value while the latter indicates there are no missing values.

num_cat_cols = num_cols+cat_cols # Combined numerical and Categorical variables

#For evert variable with a missing value create another variables with VariableName_NA

Use the numbers 1 and 0 to flag missing values

for var in num_cat_cols:

    if fullData[var].isnull().any()==True:

        fullData[var+'_NA']=fullData[var].isnull()*1

Step Five: Remove Missing Values

#We will replace or impute a missing numerical value with the mean

fullData[num_cols]

OceanofPDF.com

= fullData[num_cols].fillna(fullData[num_cols].mean(),inplace=True)

#We will replace or impute any missing categorical value with -9999

fullData[cat_cols] = fullData[cat_cols].fillna(value = -9999)

Step Six: Create Label Encoders for Categorical Variables

When you perform this step, you can split the data set into training and testing data sets. You can split the training data set further into the training and validation data set.

#We will create a label encoder to identify categorical features

for var in cat_cols:

number = LabelEncoder()

fullData[var] = number.fit_transform(fullData[var].astype('str'))

We convert the target variable since it is categorical.

fullData["Account.Status"] = number.fit_transform(fullData["Account.Status"].astype('str'))

train=fullData[fullData['Type']=='Train']

test=fullData[fullData['Type']=='Test']

train['is_train'] = np.random.uniform(0, 1, len(train)) <= .75

OceanofPDF.com

Train,Validate=train[train['is_train']==True], train[train['is_train']==False]

Step Seven: Remove Missing Values by Substituting the Dummy Variables

We will use the random forest model to predict the type of data.

features=list(set(list(fullData.columns))-set(ID_col)-set(target_col)-set(other_col))

x_train = Train[list(features)].values

y_train = Train["Account.Status"].values

x_validate = Validate[list(features)].values

y_validate = Validate["Account.Status"].values

x_test=test[list(features)].values

random.seed(100)

rf = RandomForestClassifier(n_estimators=1000)

rf.fit(x_train, y_train)

Step Eight: Check the Model's Performance and Make Predictions Based on the Analysis Results

Now that the model is ready, you can check the performance and make the necessary predictions.

OceanofPDF.com

```
status = rf.predict_proba(x_validate)

fpr, tpr, _ = roc_curve(y_validate, status[:,1])

roc_auc = auc(fpr, tpr)

print roc_auc

final_status = rf.predict_proba(x_test)

test["Account.Status"]=final_status[:,1]

test.to_csv('C:/Users/Analytics
Vidhya/Desktop/model_output.csv',columns=['REF_NO','Account.Status'])
```

You can now use this model to predict any outcomes.

OceanofPDF.com

Chapter Eleven: Data Visualization

It is important to note that people absorb information better when it is represented in a visual format. When you have performed the analysis and obtained the results, you need to find a way to interpret them. This is one of the biggest challenges that people have faced since there is a lot of information, which can be visualized. If the data is not visualized accurately, it can lead to incorrect interpretations that can impact business decisions. Therefore, as a data scientist or analyst, you need to know exactly what you are looking for in the data set and use different visual aids to depict the data.

Unfortunately, people cannot retain information in the form of text for long periods. Our brains can retain large volumes of information if they are in the visual format. Therefore, it is important to represent information in a visual format to ensure business stakeholders know exactly the analysis's objective. You can use different formats, such as pie charts, maps, spreadsheets, pivot tables, etc., to depict the information in an easy-to-understand manner.

You can convey critical concepts and aspects of the data using intuitive, swift, and simple visualization tools and techniques. Visualization also helps analysts work with different forms of data.

OceanofPDF.com

They can visualize how the data functions in different formats or scenarios that enable them to develop a robust business model. They can also make the necessary adjustments to the data set.

Data visualization is essential for organizations since it reduces the time spent in meeting rooms by at least 30%. When you represent text and numbers in the form of charts and tables, people can easily determine what you are trying to tell them. They may only have questions about how you developed the graph or table. Now that you know the importance of data visualization, let us look at some aspects to keep in mind when you develop visualization for your business.

Know Your Audience

Most people forget to understand that the objective of any visualization is to make sure the viewer understands the most important data visualization concepts. Therefore, it is important to depict the data in a format that appeals to the users. You need to find a way to tailor the visualization to suit the audience to whom you are presenting the data.

Set Your Goals

When you create the visualization for the data analysis results, you need to ensure the visuals and aids you use show logical insights and narratives relevant to the business problem being discussed.

OceanofPDF.com

When you create a goal for your campaigns and pursuits, you need to sit with the business stakeholders and explain your goals and objectives to them.

Choosing the Right Charts to Represent Data

You need to choose the right type of chart to represent the data set's data points. This form of visualization plays an important role since it is the only way you can effectively represent the data. You need to keep the purpose of the project, audience, and project in concern. For instance, if the project or query you are working on looks for the business's changes over a few months, using a simple bar graph or line graph to represent the data. These are some of the best ways to represent data visually. The following are some visualization tools and techniques you can use to represent your data set.

Number Charts

Using number charts, you can effectively and efficiently depict information, which is in numeric formats. You can use number charts to depict the number of times a user entered a website, the type of pictures a user likes on Instagram, and the organization's sales.

Maps

One of the advantages of using maps to represent the data graphically is that they are fun to look at. This means the audience will always be engaged when they look at maps. Another advantage of using maps is you can easily represent information in the data set.

OceanofPDF.com

You can also represent large volumes of complex information regarding different topics using maps. It is easy to digest the information depicted through maps.

Pie Charts

Pie charts are used as a traditional way to represent information. Most people do not like using pie charts to represent data, and these charts have received negative feedback over the years. Pie charts are still a great way to visualize data in the data set.

Gauge Charts

You can represent data with single data points or values using gauge charts. These charts are the best forms of visual representation that display trends and indicate any hidden correlations in the data set. You can also use them on executive dashboard reports shared with stakeholders.

These are not the only examples of data visualization techniques. There are other complex forms of visualization that you can use to ensure the business and stakeholders understand the analysis performed.

OceanofPDF.com

Conclusion

If you are looking to work with large volumes of data, you have come to the right place. This book has all the information you need to know about data mining, analysis, and visualization. Data mining is the process of extracting and cleaning data collected from different sources that will help you interpret and analyze the information to make informed business decisions.

The book also talks about data analysis and its importance. You will also learn more about how you can use Python to manipulate and work with different data sets. The book also leaves you with an example of how you can develop a predictive model using Python. You can use the book's code to help you learn more about how you can develop different models to work with large volumes of data.

If you are keen on working in the field of data analysis or science, you need to learn how to code in Python. Most professionals use this language to develop robust models. You also need to understand the importance of data visualization, and this book throws some light on this topic. It leaves you with some quick data visualization techniques that you can use to let the business know what you are trying to convey through your analysis.

OceanofPDF.com

References

Data Cleaning in Python. (2021). Dezyre.com. https://www.dezyre.com/article/data-cleaning-in-python/406

Data Mining Architecture - Javatpoint. (n.d.). Www.Javatpoint.com. https://www.javatpoint.com/data-mining-architecture

Data Mining Techniques - Javatpoint. (n.d.). Www.Javatpoint.com. https://www.javatpoint.com/data-mining-techniques

Data Mining Tutorial - Javatpoint. (n.d.). Www.Javatpoint.com. https://www.javatpoint.com/data-mining

Data Mining Tools - Javatpoint. (n.d.). Www.Javatpoint.com. https://www.javatpoint.com/data-mining-tools

Rodriguez, M. (2020, July 29). How to Summarize Data with Pandas. Medium. https://medium.com/analytics-vidhya/how-to-summarize-data-with-pandas-2c9edffafbaf

What is Data Analysis? Types, Process, Methods, Techniques. (2019, September 7). Guru99.com. https://www.guru99.com/what-is-data-analysis.html

What Is Data Mining? (2019). Oracle.com. https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON002

OceanofPDF.com

Data Science for Beginners

Comprehensive Guide to Most Important Basics in Data Science

Alex Campbell

OceanofPDF.com

Introduction

In today's world, organizations and people focus more on big data and artificial intelligence. It may seem surprising that over 2.5 exabytes are created and collected by people and organizations each day. This means the volume of data has risen significantly over the years. Most companies have changed their business model and centered it more on data. Some organizations have also added new departments in the firm to perform data analysis. Statisticians would need to analyze the data quantitatively in the past, but this is not enough since the results of the analysis could only talk about the present. When strong computing processes, cloud technology and analytical tools came into existence; people began to use them to perform analysis. They began to develop models to analyze data.

Let us first understand what data science is before we delve into various aspects of data science. In simple terms, data science is a branch of mathematics and statistics to obtain useful and meaningful insights about the data set and trends from the raw data or information. You can process and manage the data set using programming, business and analytical skills. This sounds tough, does it not? Most people do not know how to work with data science or understand how to develop skills effectively.

This book deals with all the information needed to learn and work with data science techniques and tools. Since the world is turning to make decisions every day, it is important to learn more about what data science is and how it helps make predictions. This book aims to give you a brief of what data science is and how you approach a problem.The field of data science goes back to its roots in statistics.

OceanofPDF.com

Having said that, this field is a combination of programming, business acumen and statistics. gIt is important to learn more about each topic, so you have an idea of how you approach the learning process. The art of finding any hidden insights and trends from the data set goes way back. Ancient Egyptians analyzed census data to help them collect tax efficiently. They also used data analysis to forecast when there could be floods in the Nile. It is important to learn from past data to identify a trend or insight in the data set. This helps the business make informed decisions.

It is no longer a secret that data scientists are in demand, and if you enjoy working with data, you should pick this field. If you learn data science, you need to grab an opportunity to work in this field. Employees skilled in data science make it easier for the company to make informed decisions. The number of data science jobs has increased over the last few decades.

A data scientist is expected to develop the necessary technical and non-technical skills. If you want to excel in this field, you need to develop these skills. The book sheds some light on these skills, which will help you identify any patterns in the data set. Using these skills, you can identify hidden patterns and insights in the data set. You can also understand the different changes they may need to make in the existing products and services to improve the business's revenue. With the information in this book, you will learn everything there is to know about data science. You will learn more about the various subjects and frameworks used in data science. So, let's get started!

OceanofPDF.com

Chapter One: Introduction to Data Science

Data has become the new oil, and every company, regardless of the industry, is looking for ways to manage and store large volumes of data. This was a challenge for most companies until 2010. Every company's objective was to define a framework or solution that allowed them to store large volumes of data. The introduction of Hadoop and other platforms has given organizations an easier way to store large volumes of data because of which they now focus on methods and solutions to process information. This can only be done using data science. It is important to note that data science is the future of technology. It is important to know what data science it, especially if you want to add some value to your business.

Introduction to Data Science

Data science is a mix of numerous algorithms, tools, principles, and languages to identify the hidden patterns within the variables in the data set. This may lead you to wonder how this is different from what has been done on data for years. The answer is that earlier we could only use tools and algorithms to explain the variables in the data set, but using data science, it becomes easier to predict the outcomes.

A data analyst uses the data only to explain what is happening in the present using historical data set. On the other hand, a data scientist only looks at the data to obtain any insights from the data set.

OceanofPDF.com

He also uses some advanced algorithms to identify the probability of the occurrence of an event. He looks at the data from various angles and aspects.

Data science is used to make informed decisions based on predictions made using the existing data set. You can apply numerous analytics to the data set to obtain this information. We will discuss these in brief in the subsequent sections.

Predictive Casual Analytics

If you want to develop a model predict the possibilities or outcomes of a futuristic event, you need to use predictive causal analytics. Let us assume you work for a credit company, and you loan people money based on their credit. You are going to be concerned with your customers' ability to repay the amount you have loaned to them. You can develop models to perform a predictive analysis on the data using the payment history. This can help you determine if the customer will pay you on time or not.

Prescriptive Analytics

You may need to use a model that can take the required decisions and modify the parameters based on the data set or question. To do this, you need to use prescriptive analytics. This form of analytics is more about providing the right information so that you can make an informed decision.
You can also use this type of analytics to predict a range of associated outcomes and prescribed actions. An example of this type of analytics is a self-driving car. We have looked at this earlier as well.

OceanofPDF.com

You can run numerous algorithms on the data collected from the cars and use the results to make the car more intelligent. This makes it easier for the car to take the right decisions to turn, slow down, speed up or identify the direction to take.

Machine Learning

Make Predictions

Numerous machine-learning algorithms allow you to make predictions using unstructured, semi-structured and structured data sets. Let us assume you work for a finance company and you have the transactional data available. You need to develop a model to determine the trend of future transactions. To perform this analysis, you need to use a supervised machine-learning algorithm. Such algorithms are used to train the machine with an existing data set. You can also use supervised machine learning algorithms to develop and train a model to detect future frauds based on historical information.

Pattern Discovery

Not every data set has variables you can use to make the necessary predictions. This is not true. There is a hidden pattern in every data set, and you need to find those patterns, so you make the required predictions.

To do this, you need to use an unsupervised model since you do not have any pre-defined labels in the data set using which you can group the variables.

One of the most common algorithms used to identify patterns is clustering.

OceanofPDF.com

Let us assume you work for a phone company, and you are tasked with identifying where to set up towers in an area to establish a network. You can then use the clustering algorithm to identify where you can set up towers to ensure every user in the area receives the optimum signal strength.

Based on the examples above, it is important to understand how data science and data analytics approaches are different. The latter includes the use of predictions and descriptive analytics only to an extent. On the other hand, data science is more about the use of machine learning and predictive casual analytics.

Now you have an idea of what data science is, let us understand why organizations need to use it in the first place.

Why Use Data Science?

Before organizations collected data from every device used, they worked with small volumes of data. It was easy to analyze and understand the data and relationships within the data set using some business intelligence tools. Most traditional business intelligence tools only worked on structured data sets, but most of the data collected today is either semi-structured or structured. It is important to understand that most data collected now are semi-structured or unstructured.

Simple business intelligence tools cannot process this type of data, especially since large volumes of data are collected from different instruments. It is for this reason we need to develop advanced and complex analytical algorithms and tools to process, analyze and draw some insights from the data.

OceanofPDF.com

It is not only for this reason why data science has gained popularity. Let us look at how data science is used in different domains.

Customer Service

How great would it be if you could know exactly what your customers want? Do you think you can use existing data to learn more about your customers, such as purchase history, browsing history, income, and age. You may have had this data with you in the past, as well. Since you use different mathematical and statistical models, you can effectively work with large volumes of data and identify the right products to recommend to your customers. This is a great way to bring more business to your firm.

Self-Driven Cars

How would you feel if your car could drive you home? Numerous companies are trying to develop and improve the workings of a self-driven car. The cars collect live information from various sensors, such as lasers, radars, and cameras. to create a map of the surrounding environment. The algorithm in the car uses this data to decide to speed up, slow down, park, stop, overtake, etc. These algorithms are often machine learning algorithms.

Predictions

Let us now consider how you can use data science in predictive analytics. Consider weather forecasting. The algorithms used take data from aircraft, satellites, radars, ships and other parts to collect and analyze data. This helps you build the required models.

OceanofPDF.com

You can use these models to predict the occurrence of any natural calamities. Using this information, you can take the necessary measures to save lives.

Who is a Data Scientist?

If you look for the words data scientist on the Internet, you may come across numerous definitions. A data scientist uses data science to answer some business questions and concerns. The term data scientist was coined when people learned that a data scientist uses data, various mathematical or statistical functions and operations, and other scientific fields and applications to make sense of the data in the database.

Functions Performed by Data Scientists

A data scientist is one who cracks various data problems using their expertise in specific scientific disciplines. He works with different mathematical, statistical, and computer science elements. He does not necessarily have to be an expert in these fields. That said, he would use some technologies and solutions to develop the right solutions and reach conclusions crucial for the organization's development and growth. A data scientist finds a way to present the data in a useful form when compared to the data available in the data set. They work with both structured and unstructured data.

Let us now look at what business intelligence is, and how it is different from data science. You may have heard about business intelligence, and most people confuse data science with business intelligence. We will look at some differences between the two, so you understand better.

OceanofPDF.com

Differences Between Data Science and Business Intelligence

Before we look at the differences between data science and business intelligence, let us understand these terms better.

Features	Business Intelligence (BI)	Data Science
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP)
Focus	Past and Present	Present and Future
Tools	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R

Using business intelligence (BI), an organization can find insight and hindsight in the existing data set to describe various trends in the data set. Through BI, businesses can take data from both internal and external sources, prepare that data, and run queries on the data set to obtain the required information. They can then create the required dashboards to answer different questions or identify solutions to various business problems. BI can also help businesses evaluate certain futuristic events.

OceanofPDF.com

Data science, on the other hand, is a different approach of looking at data. You can take a forward-looking approach and explain any information or insight in the data set. Using data science, you can analyze the current or past data that helps you predict the outcomes. This is one way most organizations do their best to make informed decisions. They can answer various open-ended questions. The following are some features that differentiate data science from business intelligence:

Now you have an idea of what data science is, let us look at the lifecycle of the data science. Most people rush into using the models they develop on the data sets without understanding the basics of data science. You need to understand these basics and assess the business requirements before you rush into using the model. Make sure to follow the phases of the data science lifecycle to ensure your results are accurate.

Lifecycle

This section gives you a brief overview of the phases in the data science lifecycle.

Phase One: Discovery

Before you work on the project, you need to understand the following:

- Business requirements
- Specifications
- Required or approved budget
- Priorities

OceanofPDF.com

If you want to pursue a career in data science, you need to possess the ability to ask important questions. You need to assess if you have the right resources, people, technology, data and time to support the work done on the project. This is the phase where you frame the problem and identify the initial hypothesis you want to test.

Phase Two: Data Preparation

When you identify the required resources needed to work on the analysis, you need to develop or identify an analytical sandbox where you can perform the testing and analysis of the data. You need to process, explore and condition the data before you model it. You also need to perform the following operations to move the data into the sandbox environment:

- Extract
- Transform
- Load
- Transform

Most data scientists use R or Python to clean transform and visualize the data used in the analysis. These programming languages help you to identify the outliers in the data. You can also use the information to develop or identify a relationship between variables. Once the data is cleaned and prepared, you can perform different types of analysis on the data. Let us look at how you can do this.

OceanofPDF.com

Phase Three: Plan the Model

During this phase, you need to identify the techniques and methods to help you draw the relationship between the different variables in the data set. These relationships will help you determine the algorithms you can use in the next phase of the lifecycle. To do this, you need to apply exploratory data analytics methods and tools using various formulae and visualization methods. Let us look at some tools used for this below:

- **R**: This programming language has various modeling capabilities. It is also a good platform to use to develop the right models if you are a beginner.
- **SQL**: SQL provides a set of methods used to perform analysis within the database using different predictive models and mining functions.
- **ACCESS or SAS**: These tools can be used to access data from various storage platforms, like Hadoop, and use that data to create a reusable and repeatable model.

The market has numerous tools you can use to develop modeling techniques, but R is commonly used. At the end of this phase, you will have the required insights in your data that will help you determine the algorithm to use. The next phase is where you apply this algorithm and develop the model.

Phase Four: Build the Model

Now that you have decided which algorithm to use, you need to split the data set into training and testing data sets.

OceanofPDF.com

In this phase, you need to consider the existing tools and determine if they are sufficient for the purpose of building a model. Make sure you identify a robust environment to run the models. You need to analyze different techniques, such as clustering, classification, and association, to develop the model. You can use different tools to build the model.

Phase Five: Operate the Model

In this phase, you run the data through the model and deliver the reports and necessary technical documents. Additionally, you may also need to run the model in the production environment to test if it works the way it needs. This gives you an idea of how the model performs on real-time data. You can also determine any constraints in the model.

Phase Six: Communicate the Results

It is important to evaluate if the model has given you the results you needed. You can do this by analyzing your hypotheses. This is the last phase of the data science lifecycle and is where you identify the key findings and communicate the same to the organization. You can determine the results of the model based on the criteria you identified in the first phase.

OceanofPDF.com

Chapter Two: Pros and Cons of Data Science

Data science is an upcoming field, and there are multiple opportunities. Having said that, this field has its share of pros and cons. This chapter looks at some pros and cons of data science, to help you take the necessary course of action.

Pros

There are many advantages to data science, and this section lists them.

Fastest Growing Field

Data science is an upcoming field and is quite in demand. If you want to become a data scientist, now is the time to do it!

Abundance of Roles

Only some people have the skills necessary to become a data scientist. You need to learn different skills and constantly learn if you want to survive in the field. This makes the field less saturated when compared to other machine learning and big data jobs. You have a lot of opportunities if you want to switch to the field of data science. There is a very low supply of data scientists.

A Versatile Field

Data science can be applied in numerous fields, but is often used in healthcare, consultancy services, e-commerce industries and banking.

OceanofPDF.com

Data science is versatile, and you have the chance to work in different fields.

Makes Data Easier to Use

Every company needs skilled employees to collect, process, analyze and visualize the data they collect. These employees are data scientists and they not only analyze the data but also improve the quality of that data. A data scientist knows what to do to improve and enrich the data that helps the company make informed decisions.

A Prestigious Career

A data scientist allows a company to make the right decisions. Many companies have hired data scientists to provide the right information they can use to make informed decisions. This gives a data scientist an important position in the organization. Since most companies are looking for data scientists, you can earn a lot of money. According to Glassdoor, you may earn close to $160,000 per annum.

Remove Redundancy

Data science is used in different industries, and most algorithms used in data science help to reduce the number of redundant tasks performed by people. Most companies collect historical data, and they can use this data to train machines to perform redundant tasks, which helps to simplify some jobs performed by people.

Products Become Smarter

Data science is a field that involves the use of machine learning.

OceanofPDF.com

There are different types of algorithms used in machine learning, known as supervised, unsupervised and reinforcement learning, which look at data sets to determine customer's behavior. For instance, most e-commerce websites use recommendation systems to provide insights to the customers based on their purchase history. This has made it easier for computers to understand how humans behave.

Save Lives

The healthcare sector uses data science to improve diagnoses and predictions made about patients. Through the use of machine learning algorithms, the healthcare sector has found a way to detect tumors and cancer at an early stage. There are many other benefits the healthcare industry gains from using data science.

Aid in Personal Growth

Data science is not only a great career, but it also helps you grow in your career and personally. If you choose to become a data scientist, you will develop the right attitude and thought process to solve problems. Since the field of data science is a mixture of management and IT, you learn from both areas of business.

Cons

Data science is a hot career, and many people choose to become a data scientist because it pays well. That said, there are some disadvantages to the field. If you want to better understand data science, you also need to look at the disadvantages of data science.

OceanofPDF.com

Data Science is an Ambiguous Term

There is no definite definition or meaning of data science. It has become a term people use often to talk about analysis, so it is hard to know what data science actually is and what a data scientist can do. The role of a data scientist is dependent on what the company does.

Impossible to Master Data Science

As mentioned earlier, data science is a mix of numerous fields, such as computer science, mathematics and statistics. It is impossible to master the fields used in data science, which means you cannot become an expert in those fields. While most online courses have been doing their best to fill the gap that people in the data science industry are facing, it is impossible to do this. People with a background in statistics may not necessarily have the necessary information in computer science. This field changes frequently, and you need to keep learning different areas of data science if you want to stay abreast.

Requires a Lot of Domain Knowledge

From what you have read in the previous chapter, you know you need a lot of domain knowledge. If you have enough knowledge about computer science, statistics, and mathematics, you may not easily solve a data science problem if you do not have background information. The same can be said for the reverse, as well. Let us assume you work for a health-care company where you need to analyze genomic sequences. To do this, you need to have some information about molecular biology and genetics. This is the only way you can make calculated decisions that help the company. If you do not have this background, it will be difficult for you to work on analyzing genomic syndromes.

Unexpected Results

Data scientists analyze the information in the data set and use the patterns and variables in the data set to make informed decisions. This helps you make informed decisions. There are times when the data provided is arbitrary and you may not obtain expected results. The results may also vary due to the poor utilization of resources and weak management of data.

Lack of Data Privacy

Data is the new oil for numerous industries, and most companies hire data scientists to use the data they collect to make informed decisions. Having said that, the data used in these processes may lead to a privacy breach. The personal data of most clients is stored by parent companies, and some companies do not have enough security to prevent any data leaks. Many countries now have issued regulations and guidelines to prevent such leaks and secure the data.

OceanofPDF.com

Chapter Three: Statistics for Data Science

This chapter sheds some light on the statistical concepts used in data science. It is important to remember that data science is not a new concept, and most statisticians can work as data scientists. Data science uses many concepts from statistics, since this is the best tool to help you process and interpret information in the data set. There is a lot of information you can gather from the data set using statistical tools. If you want to master data science and become an expert in the field, you need to learn everything you can about statistics. While there are many topics in statistics a data scientist should know about, the following are the most important aspects to consider:

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics

Descriptive statistics is the process of expressing or looking at the data to help you read it easily. This method helps you deal with quantitative summarization of the data through numerical representation or graphs. Some topics you need to learn about are listed below:

- Normal Distribution
- Central Tendency
- Variability

OceanofPDF.com

- Kurtosis

Normal Distribution

A normal distribution, also known as a Gaussian distribution, is a continuous distribution often used in statistics. Any data set following a normal distribution is spread across a graph, also a bell-shaped curve. In normal distributions, the data points in the set peak at the center of the bell-shaped curve, which represents the center of the data set. When the data moves away from the mean, it will fall to the end of the curve. You need to ensure the data you look at is distributed normally if you want to make inferences from the data set.

Central Tendency

Measures of central tendency help you determine the center values of the data set. There are three measures commonly used – mean, median and mode. The mean or the arithmetic mean of any distribution is found at the center of the data set. You can use the following formula to calculate the mean or average of the data set: (sum of all points in the data set) / (number of data points)

The median is another measure that is the mid value of the data set when the points have been sorted in an ascending order. If you have an odd set of values, you can easily find the mid value, but if you have an even number of data points, you take the average of the two data points in the middle of the data set. The last measure is the mode, and this value is the data point, which occurs the maximum number of times in the data set.

OceanofPDF.com

Variability

Variability is a factor which helps you determine the distance between the data points and the average or mean of the data points in the data set. This value also shows the difference between the data points selected. You can view and assess the variability based on the measures of central measure, such as range, variation and standard deviation. The range is a value, which is the difference between the smallest and largest value in the data set.

Kurtosis and Skewness

The skewness of the data set helps you determine how symmetrical the data set is. If the data set is distributed uniformly, it will take the shape of a bell curve. If the curve is shaped evenly, the data is not skewed. If the curve moves either to the right or left side of the data points, it means the data is negatively or positively skewed, respectively. This means the data is either dominant on the left or right side of the measures of central tendency. Kurtosis is a measure that helps you determine the tails of the distribution. When you plot the data points on a graph, you can determine if the data is light or heavy-tailed. You can make this assumption based on the middle section of the distribution.

Inferential Statistics

Descriptive statistics provide information about the data, but inferential statistics are about obtaining insights about the data set.

OceanofPDF.com

Inferential statistics is about concluding about the large population based on a small data set or sample. Let us assume you need to count the number of people in Africa who have received the polio vaccine. You can perform this analysis in two different ways:

- Ask every individual in Africa if they have been given the polio vaccine
- Take a sample of people from the continent, make sure it is from different parts of the continent, and extrapolate the findings across the entire continent

The first process is difficult to complete, and it is impossible to do it. You cannot go around the entire country and ask people to tell you if they have received the vaccine. The second method is the better way to do this since you can draw insights or conclusions from the sample selected and extrapolate the findings to the large population. The following are some tools used in inferential statistics.

Central Limit Theorem

According to the central limit theorem, "the average of the sample is same as that of the entire population." This shows the properties and measures of the data's central tendency, such as standard deviation, are the same for both the sample and population. This means you can increase the number of data points selected, resulting in a normal curve.

You need to understand the concept of confidence intervals if you want to use the central limit theorem. This goes to show the approximate value of the mean of the population.

OceanofPDF.com

The process of creating an interval in the population has things like the sum of an error margin. You can calculate this error by "multiplying the standard error of mean with the z-score of the percentage of confidence level."

Hypothesis Testing

Hypothesis testing is the extent to which you can test any assumption you make about the data set. In this process of testing, you can collect the results of your analysis of the hypothesis on a smaller population. The hypothesis you need to test is called the null hypothesis, and we need to check the validity of this hypothesis against the alternative hypothesis. The null hypothesis is the case you need to test. Consider the following example – you are conducting a survey to determine who smokes and who does not, and how people who smoke have cancer.

When you conduct this survey, you begin with the assumption that the number of subjects with cancer who smoke is the same as the number of subjects with cancer do not smoke. This is your null hypothesis, and you need to test it to reject the hypothesis. The alternate hypothesis will be that the number of subjects with cancer who smoke is greater than the number of subjects with cancer who do not smoke.
Based on the given data and evidence, you can test the hypotheses and analyze the data to determine if the null hypothesis is correct or not.

ANOVA

ANOVA is another statistical concept that is used to test hypotheses across different groups of data.

OceanofPDF.com

This concept helps you determine if the groups you are checking have similar averages and variances. ANOVA allows you to carry out this form of analysis with limited error rates. You can use the F-ratio to calculate ANOVA. The F-ratio is used to calculate the ratio of the mean square error between groups to the mean square error in specific groups. The following are the methods to calculate ANOVA:

1. Write the hypotheses and understand the need for those. Every analysis should have a null and alternative hypothesis
2. In case of null hypothesis, you need to assume the average of the groups is identical
3. The alternative hypothesis will have a different average

OceanofPDF.com

Chapter Four: Skills Required for Data Science

If you want to work in the field of data science, or become a data scientist, you need to have some information about different topics mentioned in this chapter. This chapter lists some technical and non-technical skills you should develop if you want to improve in this field.

Technical Skills

As a data scientist, you need to have some technical skills that help you with statistical analysis. You need to learn how to leverage and work with different frameworks and software to mine, collect, process, collate, analyze, interpret and visualize large volumes of data. You need to develop programming skills to perform such activities. An easy way to do this is to ensure you have the necessary academic background. Most data scientists have a Ph.D. or master's degree in engineering, statistics, and computer science. This is the only way to determine if they have the foundation to help them connect with different technical points that are the foundation of the practice. Many schools offer such programs to help people pursue data science.

If you do not want to go through hardcore courses, you can look at options, such as:

- Boot camps
- Massive Open Online courses (MOOC)

These programs help you develop a basic understanding of core data science subjects.

OceanofPDF.com

These courses also provide some information that is outside of textbook learning. You will be given real-time scenarios and asked to develop models to assess and predict futuristic events. The following are some skills you need to develop.

Understanding Data

Data science is about working with and understanding different types of data. You need to understand and love working with data. The following are some questions you can answer to help you understand whether you love data or not:

- Do you know where you need to collect data from?
- Do you understand and know what data is?
- Have you determined what information you should look at?
- How frequently do you work with data?
- Have you learned how to work with structured and unstructured data?

The most important question you need to answer is whether you love working with data or not. If yes, you need to obtain certifications, so you develop the skills to become a data scientist.

Algorithms

Algorithms are sets of instructions that you write. You can use algorithms to instruct a machine to perform specific functions and tasks. Let us try to write an algorithm using that we will instruct a computer to add two numbers.

1. Identify two variables and declare it to the machine
2. Initialize the variables

OceanofPDF.com

3. Ask the user to assign a value to each of these variables
4. Declare another variable to hold the sum of the first two variables
5. Calculate the sum of the first two variables and assign that value to the third variable

You can also use algorithms when you solve puzzles on paper. As a data scientist, you need to understand what algorithms are and how a machine understands them since you work with algorithms to help you analyze data. As a data scientist, you also need to learn how to design different algorithms that perform the necessary functions to help you analyze data. Let us assume you need to key in 10 numbers into the system. You may enter any ten numbers and leave it to the machine to identify the set's third largest number. To do this, you should write an algorithm to help the machine identify the number. As a data scientist, you need to write the necessary logic and develop an algorithm which helps you find the third largest number.

Programming

It is important to learn different languages, such as Java, C++, R, Python, Perl, SQL and other languages. The languages used commonly are R and Python. You can collect, clean, process, organize and analyze the information in the data set which helps you work with unstructured data.

Examples

Example 1

This takes care of only variables and constants – Program to build Hello World

```
> # We can use the print() function
```

OceanofPDF.com

```
> print("Hello World!")
[1] "Hello World!"
> # Quotes can be suppressed in the output
> print("Hello World!", quote = FALSE)
[1] Hello World!
> # If there is more than 1 item, we can concatenate using paste()
> print(paste("How","are","you?"))
[1] "How are you?"
```

In the program above, we have used print(), a built-in function to print the required string Hello World! The quotes that you see are printed by default. To avoid that, we can add an argument called quote = FALSE. Also, if there is more than a single item, you can use paste() or cat() functions to concatenate the strings together.

Example 2

We can add the elements of the vector by the function sum()

```
> sum(2,7,5)
[1] 14
> x
[1]  2 NA  3  1  4
> sum(x)    # if any element is NA or NaN, result is NA or NaN
[1] NA
> sum(x, na.rm=TRUE)    # this way we can ignore NA and NaN values
[1] 10
> mean(x, na.rm=TRUE)
[1] 2.5
> prod(x, na.rm=TRUE)
[1] 24
```

When a vector has NA (not applicable), or NaN (not a number),

OceanofPDF.com

the functions that are used here such as sum(), mean(), prod(), etc make NA or NaN, respectively.

Example 3

This example will deal with an interactive screen, i.e. take inputs from the user.

```
my.name <- readline (prompt="Enter name: ")
my.age <- readline (prompt="Enter age: ")
# convert character into integer
my.age <- as.integer(my.age)
print (paste ("Hi,", my.name, "next year you will be", my.age+1, "years old."))
```

Output:

```
Enter name: Mary
Enter age: 17
[1] "Hi, Mary next year you will be 18 years old."
```

As you can see, we have used the function readline() to get input from the user.

Here, you can see that you can display an appropriate message for the user with the prompt argument.

In the above example, you convert the input age, which is a character vector into integer by the function as.integer().

This is necessary for doing further calculations

Example 4

In this example, we will find if a year is leap year or not by taking inputs from the user.

```
# Program to check if the input year is a leap year or not
year = as.integer(readline(prompt="Enter a year: "))
if((year %% 4) == 0) {
if((year %% 100) == 0) {
if((year %% 400) == 0) {
print (paste (year,"is a leap year"))
} else {
Print (paste (year,"is not a leap year"))
}
```

```
} else {
Print (paste (year,"is a leap year"))
}
} else {
Print (paste (year,"is not a leap year"))
}
```

Output 1:

```
Enter a year: 1900
[1] "1900 is not a leap year"
```

Output 2:

```
Enter a year: 2000
[1] "2000 is a leap year"
```

Here we have used the logic that a leap year is exactly divisible by 4 except for the years ending with 00. The century year is a leap year only if it is perfectly divisible by 400.
Nested if else is used to implement the logic in the above program.

Example 5

In this example, we will find the HCF of two numbers

```
# Program to find the H.C.F of two input number
# define a function
hcf <- function (x, y) {
# choose the smaller number
if(x > y) {
smaller = y
} else {
smaller = x
}
for(i in 1:smaller) {
```

OceanofPDF.com

```
if((x %% i == 0) && (y %% i == 0)) {
hcf = i
}
}
return(hcf)
}
# take input from the user
num1 = as.integer (readline(prompt = "Enter first number: "))
num2 = as.integer (readline(prompt = "Enter second number: "))
print (paste ("The H.C.F. of", num1,"and", num2,"is", hcf(num1, num2)))
```

Output:

```
Enter first number: 72
Enter second number: 120
[1] "The H.C.F. of 72 and 120 is 24"
```

This program asks the user to input two integers and the pass them to a function which give the output as the H.C.F.

The function first determines the smaller of the two numbers given as an input since the H.C.F can only be less than or equal to the smallest number among the two.

We then use a 'for' loop to go from 1 to that smaller number.

In each loop we check if our number perfectly divides both the input numbers.

If yes, then we store the number as H.C.F. At the completion of the loop we will end up with the largest number that perfectly divides both the numbers.

Example 6

In this example, we will show you how to develop a calculator of your own.

OceanofPDF.com

```
# Program makes a simple calculator that can add, subtract, multiply and divide using functions
add <- function(x, y) {
return(x + y)
}
subtract <- function(x, y) {
return(x - y)
}
multiply <- function(x, y) {
return(x * y)
}
divide <- function(x, y) {
return(x / y)
}
# take input from the user
print ("Select operation.")
print ("1.Add")
print ("2.Subtract")
print ("3.Multiply")
print ("4.Divide")
choice = as.integer (readline (prompt="Enter choice[1/2/3/4]: "))
num1 = as.integer (readline (prompt="Enter first number: "))
num2 = as.integer (readline (prompt="Enter second number: "))
operator <- switch(choice,"+","-","*","/")
result <- switch (choice, add(num1, num2), subtract(num1, num2), multiply(num1, num2), divide(num1, num2))
print (paste (num1, operator, num2, "=", result))
```

Output:

```
[1] "Select operation."
[1] "1.Add"
```

OceanofPDF.com

```
[1] "2.Subtract"
[1] "3.Multiply"
[1] "4.Divide"
Enter choice[1/2/3/4]: 4
Enter first number: 20
Enter second number: 4
[1] "20 / 4 = 5"
```

In this code, we will first ask the user what operation he wants to carry out. Then, the user is asked to type in two numbers, and we use 'switch' branching is used to carry out a particular function.
The functions such as add(), subtract(), multiply() and divide() areal user-defined functions.

Analytical Tools

Every data scientist needs to learn how to use different analytical tools and understand the use of Software as a Service (SaaS). This understanding will help you obtain some information from the data you have cleaned and processed using different programming languages and tools, such as Hadoop, R, SAS, Spark, Hive and Pig. Different platforms allow you to improve your skills, and you can obtain certifications that will help you move ahead in your career.

Using Unstructured Data

Data scientists need to collect, process, store, manage, clean, understand and analyze the data collected from various sources. Most of these data are not structured, and they can be collected in different forms, such as images, videos, text, emails and other forms.

OceanofPDF.com

For instance, if you work with a credit company, you need to learn how to use historical data and patterns to assist the company in identifying the customers they can trust. You also need to learn how to collect, process and analyze the data.

Non-Technical Skills

The following are some non-technical skills you need to develop if you want to be a data scientist. These are some personal skills a user needs to develop, and these are very different from qualifications and certifications.

Strong Business Acumen

As a data scientist, you need to have strong business acumen. You need to understand various elements important to the business model. Otherwise, you cannot use the necessary tools and skills to analyze the data set. It is only when you learn and develop these skills that you can identify the various problems and solutions to develop if you want to help the business make informed decisions so that it can sustain and grow. It becomes extremely hard for you to help an organization explore various opportunities if you do not know how to code.

Communication Skills

Data scientists can understand data better than anybody else in the organization. If you want to succeed as a data scientist, you need to learn to communicate your decisions and analysis to the stakeholders. This is the only way you can make it easier for the organization to make informed decisions. It is important to learn to communicate your findings to people in the organization, especially if they are non-technical.

OceanofPDF.com

You need to develop the necessary communication skills to communicate your findings.

Intuition

This is an important non-technical skill you need to develop as a data scientist. It is true you may already have an insight about the data when you look at it, but you may not have the ability to perceive any hidden patterns in the data set. You need to know where to look and what to look at, so you add more value to the analysis.

These are the skills that make you efficient in your work. Do not worry about cramming too much information in a short time. You learn and improve as you gain experience.

OceanofPDF.com

Chapter Five: Tools Required for Data Science

We have looked at different aspects of data science and know why you need to use data science. Let us now look at the different tools you need to work with to become a data scientist. The following are only some tools you need to use if you want to work in data science.

R Programming Language

We have referred to R earlier in the book. This is a programming language widely used by data scientists, and many organizations use this language to perform data analysis. Organizations use R widely since it is an object-oriented programming language, and it is used for the analysis of reporting, visual graphic representations and statistics.

Python Programming Language

Python, like R, is another programming language that has features similar to other object-oriented programming languages. It is used widely in data science and to develop models. Python can be used on different operating systems, and you can develop numerous applications using this language. This programming language is used to develop machine learning models and applications that use artificial intelligence.

OceanofPDF.com

Hadoop

As mentioned earlier, Hadoop came into picture when it became difficult for companies to store large volumes of data. This is an open-source application which allows you to store and process big data. You can use a distributed computing framework to analyze the data. It is best to use this application if you want to store data without having to process it. You can store as much data as needed before you decide which variable you need to use.

Structured Query Language (SQL)

SQL or Structured Query Language is used for data storage. You can create new tables and databases in SQL, and the only thing you need to do is update the records in the database. You can retrieve, modify and delete the records in SQL databases and tables. You can use this language to manage the data stored in a database.

SAS

SAS or Software as Service is an important tool to understand if you wish to work as a data scientist. This tool is designed to perform various statistical and mathematical operations. Large organizations use SAS to identify or draw hidden patterns within the data set. SAS uses SAS coding language that is often used by statisticians to develop models.

OceanofPDF.com

The language was developed by statisticians and professionals and is based on Reliable Commercial Software (RCS). This programming language has many statistical tools and libraries that data scientists can use to perform modeling. They can also use this language to keep the records in order. SAS is not an open-source programming language. This language is costly and can only be used by large organizations. There are some packages and libraries in SAS that cannot be accessed by organizations in the base pack, and you need to spend money to update the language.

Apache Spark

Apache Spark is a popular analytics tool used by data scientists, and it is designed for stream processing and batch processing. There are numerous APIs developed by data scientists across the globe to develop a frequent path for data to flow through for any machine learning algorithm. This application is a win when compared to Hadoop and other data processing and storing applications since it works faster. It comes with numerous APIs for machine learning and deep learning which enables data scientists to obtain hidden insights in the data set. The application can stream large volumes of data and is more efficient when compared to other big data platforms. This tool can work with real-time data, and allows you to code in Java, R and Python. You can use this application when you work with clusters of data. This tool is more efficient when compared to platforms, such as Hadoop and MapReduce.

OceanofPDF.com

D3.js

This is a Javascript library, and you can use this to develop visualization on the Internet. There are numerous APIs of this library, and you can use various functions to make visualizations more dynamic. You can also analyze the data set collected on the Internet. One of the advantages of using this tool is it allows you to use animated transitions to improve data visualizations. When you use this tool, you can create dynamic documents and use options to change the data used in the visualization.

You can also use this tool along with CSS to create transitory and eminent visualizations that would show you the required graphs on the Internet. This is an important tool to use in data science, and you can work with different IoT devices. The tool allows you to work with client-side interventions for data visualization and processing.

OceanofPDF.com

Chapter Six: Application of Data Science

Now that we know the importance of data science and the prerequisites and skills needed for data science, it is important to know how data science can be applied in the real world.

Risk and Fraud Detection

Data science was first used in the financial and banking sector. Many financial institutions were sick of bad debts and their losses at the end of each year. Since they had collected large volumes of data when the initial paperwork is carried out, they had a wealth of information they could use to decide who they could give the credit to. They hired a data scientist to prevent future losses caused by bad debts. Financial institutions now have learned to collect, process, divide the data and assess it. They divide the data based on customer profiling. It also helped in the sales of banking products as they could do targeted marketing based on how much a customer earned every year.

Healthcare

Many healthcare companies use medical image analysis, content-based medical image indexing, wavelet analysis and machine learning models to detect tumors, organ failure and other procedures and frameworks. These processes make it easier to analyze and find the optimal parameters or points to improve diagnoses.

OceanofPDF.com

Genetics and Genomics

You can use data science to personalize treatment for patients, improving the treatment administered to patients for various illnesses based on research in genomics and genetics. This process aims to determine what the consequences of DNA are on our health and determine if there are any connections between genes and diseases.

Drug Development

It takes 12 years to develop a drug and a few months before it is tested and becomes an official drug. The process of identifying the drug and assessing if it will work on people is complicated and involves the use of many disciplines. Billions of tests often restrict the greatest ideas. You can use machine learning and data science to simplify the process and reduce the time taken to initially screen the drug compounds used to develop the drug. Algorithms and data science can also predict how the body will react to certain compounds in the drug using various statistical and mathematical simulations and models. This is faster to when compared to traditional lab experiments. The models can predict future outcomes with better accuracy, thereby simplifying the process.

Internet Search

When you think of data science, the first thing that comes to your mind is the search algorithm used on the Internet.

OceanofPDF.com

Every search engine uses various data signs to provide the user with the best results of any query you key into the search tab. Google would not be the search engine it is today if the developers did not use data science algorithms to improve search results.

Website Recommendations

Most websites, such as Netflix, Amazon, IMDB and Twitter, use historical data to send suggestions to the users or customers to improve their experience. For example, if you want to make a purchase of earrings or clothes online, you may look at one website. Google will use that historical search to send you more suggestions. Instagram uses a similar profile as well. The recommendations made by these platforms are based on previous searches you may have performed.

Every marketing company makes use of data science algorithms to send suggestions to customers. Every advertisement you see at the airport or theatre on a digital billboard or screen on the website uses data science algorithms. Digital advertisements have been doing as well as they have since they can get a higher click-through rate when compared to traditional advertisements. You can use data science algorithms to use historical data based on the past behavior.

Advanced Image Recognition

You may have uploaded a picture on Instagram, Facebook or other social media platforms.

OceanofPDF.com

These platforms may send you a suggestion to tag your friends in the picture based on your previous pictures. This feature uses the face recognition algorithm to suggest whom you can tag. Google also allows you to search for some images by uploading specific images. It provides related search results by using image recognition.

Virtual Assistance

Many organizations have developed a mobile application that helps patients visit doctors virtually to determine what is wrong with them. These applications are powered with artificial intelligence and have a chatbot that allows them to speak to healthcare professionals to understand their symptoms. There are many applications you can use to help remind you to take medication on time. Another example is the use of chatbots on different websites. You can speak to the chatbot and ask them any question you may have.

Speech Recognition

Siri, Cortana, Google voice and other voice assistant products use speech recognition to improve your experience with different gadgets. Your life may stop if you are not in a position to type. These tools convert the words you say to text, but it is important to note that this method of conversion does not perform accurately.

OceanofPDF.com

Planning Routes for Airplanes

Airline companies have begun to use data science to identify different routes to optimize the airline routes. To reduce the heavy losses, airline companies bear due to an increase in fuel prices and heavy discounts to customers due to high competition, companies can use data science for the following:

- Predict delays in flights
- Decide the type of aircraft to use or buy
- Determine which flight you should suggest to a customer without any halt or stop
- Developing customer loyalty programs

Gaming

Most developers have started using machine-learning algorithms in games. These algorithms learn from how the user plays the game and uses that information to help the player upgrade or move to higher levels. Even in motion gaming, the computer uses previous or historical data and analyzes it to provide the best experience. EA Sports, Sony, Zynga and Nintendo have taken the experience of gaming to a whole new level by using various concepts of data science in their code.

Augmented Reality

This is an interesting and exciting application of data science,

OceanofPDF.com

and it is something many companies are trying to incorporate in their products. If you do not know what I am talking about, think about virtual reality headsets and glasses. These products contain computing algorithms, logic and knowledge to improve your viewing experience.

Data science and virtual reality have always had a strong relationship. Virtual reality products allow you to experience things you could not have in the past. For example, in the show' Bones,' Angela Montenegro, an employee of the Jeffersonian institution, sets up a virtual reality monitor that allows a child to take a trip in the Louvre. There are similar activities in Disney Land, United States, where you can take a parasailing trip without actually being in air.

You may relate to the use of data science in virtual reality when it is easier to price the value.

OceanofPDF.com

Chapter Seven: Algorithms and Statistics

Do you need to become a Ph.D. in mathematics or statistics to become a data scientist? No, you do not. The extent to which you need to know depends on what role you want to take up in the organization. This chapter gives you an idea of all the concepts you need to know if you want to become a data scientist. As a data scientist, you first need to understand the basic concepts of probability theory and statistics. If you want to master different statistics and mathematics algorithms, you need to practice them regularly. You need to follow a top-down approach if you want to learn how to code. You need to learn how to use a data stack, familiarize yourself with real-world projects and use libraries and documentation. When you notice your lack of theoretical background, you can see the bigger picture. It is important to learn how these algorithms work. Start off with the mathematical aspects before you move onto those areas, which are more complex. This chapter has the list of statistical and mathematical concepts you need to learn.

Naïve Bayes

A Naïve Bayes classifier is an algorithm which uses the principles of probability. One of the most important principles it uses is that the value of one feature in the data set is independent from the value of other features in the data set.

According to this theorem, you can predict the probability that an event will occur based on the conditions of the occurrence of that event.

OceanofPDF.com

It is important to learn how these classifiers work, and to do this, you need to study the basics and principles of conditional probability and probability. One of the easiest ways to understand this theorem is to study the concepts of probability. You can then check on Bayes' theorem and how to code this theorem.

Linear Regression

Linear regression is a basic form of regression, and it allows you to determine the connection between two variables in the data set. One of the variables is termed as the predictor variable and the other is known as the response variable. These variables are often continuous in nature. A simple linear regression model uses various data point sets and then extrapolates the trend of the values and how they will progress in the future. Linear regression is a common algorithm used in parametric machine learning, and in this form of learning, you can train the machine to develop a mathematical function using existing data sets. You can use this function to make predictions or forecasts about future events. These functions are termed as models in machine learning terminology. The best way to learn more about regression is to develop an understanding of elementary statistics. If you want to learn more about regression and the various concepts, you need to take an advanced course of statistics.

Logistic Regression

Logistic regression is a process sued in different cases where you have a binary dependent variable.

OceanofPDF.com

This form of regression focuses on the estimation of probability of the occurrence of an event. Logistic regression is also an algorithm used in parametric machine learning. Like linear regression, this algorithm also results in a mathematical function. The difference between these forms of regression is that logistic regression leads to developing a mathematical function that estimates the values based on their logarithms or other statistical functions. Another difference between linear and logistic regression is that the former uses real numbers and gives those as the output. The latter uses a model which provides the logistic function. A logistics function produced is also termed as a sigmoid function that takes care of the values to give you an output between 0 and 1. Let us understand why a sigmoid function returns a probability value. It is because of the algebraic concepts used in the algorithms that a negative exponent is taken.

Neural Networks

A neural network is a form of machine learning algorithm. These forms of networks are based on the structure of the neurons in the brain. The neural network uses a series of activation points or weights to help you make the necessary predictions. The neurons take the input, apply the transformation function and give you an output that helps you make the necessary decisions. When it comes to capturing any nonlinear relationships within the data set, it is best to use a neural network best since it helps in tasks, such as image and audio processing. The fundamental concept of neural network is to transform the input and processing it to generate the output.

OceanofPDF.com

If you want to understand the math better in a neural network, you need to take up courses to help you learn more about linear algebra and geometry. This is a better way to start preparing. If you want to delve deeper, you need to learn more about graph theory, matrix theory and real analysis.

K-Means Clustering

The k-means clustering algorithm is a type of unsupervised machine learning algorithm. You can use this algorithm to categorize any unlabeled data. This type of data does not have any defined categories or groups, but it works to identify or define any hidden categories in the data set. The number of variables identified by the algorithm are 'k' in number. The algorithm iterates over the various data points in the data set, so every point is assigned to one of many groups. This algorithm depends on the idea that the distance between the data points and the center of the cluster is the same across all data points. You can use any function that shows the distance between the elements in the data set. Suppose you want to work with this algorithm. In that case, you need to understand the basics of algebra, including addition and subtraction to identify the distance between the data points easily. If you want to delve deeper, you need to learn about Euclidean and non-Euclidean geometry.

Decision Tree

Decision trees are an easy way to determine the outcome of any decision you take using a flowchart. The flowchart is in the form of a tree structure that uses a branching method.

OceanofPDF.com

Every node in the decision tree is based on a specific variable and every branch is an image of the outcome. A decision tree is based on information theory, since this is how they are constructed. An important concept of a decision tree is known as entropy. Entropy is a measure of uncertainty in any variable. When you use entropy, you can construct a decision tree easily. When there is higher entropy, there is more uncertainty in the data. this means you need to split the tree in a way to decrease any uncertainty.

Information gain helps you determine how much information you can gain from someone. When it comes to a decision tree, you can calculate the information gain on every column based on previous columns based on the information present in the columns in the data set.

Here is a small tip – you need to learn basic algebra and probability to know what decision trees are and how they work. If you want to learn more about decision trees, you need to learn about logarithms and probability.

Some Final Thoughts

Mathematics and statistics are difficult subjects, and they can feel daunting and dry at times. You will, however, feel equipped when you compare your skills against the skills of your peers. It is also important to learn how you can apply these concepts on various data science problems.

The topics highlighted in this book will need to be read well and understood. This is the only way you can develop and build algorithms.

OceanofPDF.com

You can rely on different machine learning libraries and tools to complete this work for you. It is useful for a data scientist to understand statistics and math, which helps you determine what is happening in the tools or algorithms. This allows you to choose the algorithm that works best for your data set and allow you to make the necessary predictions.

Dive deep and work hard to develop the skills necessary to become a data scientist.

Applying Math to Data Science Models

If you want to work as a data scientist, you need to have some mathematical skills. This is the only way you can understand data and the significance of the patterns in the data set. These are important data science skills since you use them to build and develop models and perform predictive analysis and hypothesis testing. Data scientists use math to develop decision models and they can use different mathematical algorithms to predict the future. This book explains some concepts you need to learn and the skills you need to develop if you want to become a data scientist.

Deriving Insights from Statistical Methods

If you are a data scientist, you need to understand the basics of statistics. You should understand the significance of data, develop hypotheses and validate them by simulating certain scenarios. This makes it easier for you to predict futuristic events. It is very difficult to develop advanced statistical skills, but if you are keen about pursuing a career in data science, you need to understand some basics of logistic regression, linear regression, time series analysis, Bayes classification, etc.

OceanofPDF.com

Some Essentials for Data Science

Understanding programming languages and coding in various languages, especially R and Python, are important for data science. You need to learn how to write code to instruct the machine on how it should process, manipulate, analyze and visualize the data. It is important to understand how to write code in R and Python. There are different libraries and functions in these languages that allow you to manipulate, analyze and visualize the analysis of the data set.

You can also use SQL to run queries on the data set that allows you to extract and modify data in the database. People also use JavaScript library to develop interactive, dynamic and custom visualizations on the Internet. To do this, you need to learn to write code in R or python. It is important to learn to code if you want to become a data scientist, and these are easy languages to learn.

OceanofPDF.com

Chapter Eight: Data Science in a Sector

Most statisticians have been stating that data science is exactly what they have been doing for the last few decades. While this may be true to an extent, you cannot agree with them. Data science is completely different and separate from traditional statistical and mathematical approaches. Data science is different from statistics in many ways. Data scientists use programming languages to improve the efficiency of their analysis. If you are a statistician, you need to know the subject in its entirety. As a data scientist, you need only to understand the basics of mathematics and statistics.

Statisticians have very limited expertise about anything apart from statistics. They need to reach out to subject matter experts if they want any help or information about other subjects. They also need some help from experts to help them determine and understand their findings. This is the only way they can learn how they can move forward.

On the other hand, a data scientist has enough information about different subjects that helps him understand and analyze the findings with ease. A data scientist generates deep insights and uses his subject matter expertise to understand what the data is trying to tell him.

The following are some ways to use data science to improve the performance of the business in specific sectors.

- Some architectural firms have begun to use machine learning algorithms to optimize the consumption of energy in modern designs

OceanofPDF.com

- Marketing firms and businesses use logistic regression and other statistical algorithms to predict how customers may behave in the future
- The healthcare industry uses data science to determine illnesses and personalize treatment plants. This allows doctors to predict and check for any future health problems
- A data journalist takes information from the website to identify any patterns and connections within the data set
- Law enforcement agencies and organizations use data science to prevent and predict any criminal activities

Communicate the Insights Captured Through Data

If you want to become a good data scientist, you need to have sharp written and oral communication skills. As a data scientist, you need to communicate the information you have inferred from the data. If you cannot do this, all the information in the data set means nothing to you or the business. A data scientist must find a way to explain the insights gathered from the data set so that he can develop meaningful and clear visualizations.

It is important to remember that people are visual, so they learn better when they can visualize anything. As a data scientist, you need to be pragmatic and creative about the data collected and find the right way to communicate them clearly.

OceanofPDF.com

Leverage Cloud-Based Solutions

If you are not used to working with large volumes of data, you can use cloud-based solutions to learn more about the patterns in the data set. Most organizations and data analysts have understood the use of data analytics and big data and their importance. It is a new domain for some people and organizations, but they are working as hard as possible to prepare for the flood of big data. many cloud applications, both private and new, are dedicated to improving the way companies handle their big data.

You can also use cloud services, such as Tableau, to clean and process data sets. Most cloud-based applications are open-source and code-free. You can use these solutions to model and visualize data using various statistical tools and techniques. You need basic knowledge of mathematics and statistics before you apply any data science techniques on the data set. If you do not have knowledge about programming, you can use applications that do not require you to write any scripts or code.

Regardless of whether you use cloud-based solutions to perform data analysis or not, you need to train your employees and gather more information about what data science is. You also need to learn about the algorithms discussed earlier. It is important for you to learn more about how to process data, design a model, build the model and analyze it if you want to obtain insights about the data.

OceanofPDF.com

You cannot expect any cloud-based platform to make it easier for anybody to work with the data set if they do not have basic knowledge of mathematics and statistics.

OceanofPDF.com

Chapter Nine: Data Cleaning Using R

Data is the new fuel for all organizations, and they collect data from every source possible. This means they work with unstructured, semi-structured and structured data. It is for this reason we need to clean and process the data, so you transform raw data into structured data that is easier to analyze. Cleaning and processing of data influences the analysis you perform. Most data scientists use R to perform an analysis on data, and this language offers a comprehensive and effective method to analyze data. Let us look at the process of cleaning the data.

Step One: Exploratory Analysis of Data

When it comes to processing information in the data set, you need to explore the data. You can import the data into R and use different operations to explore the data set. Before we perform the analysis, it is important for you to understand how to import the data into R.

```
setwd("C:/Users/USER/Desktop/House Pricing ")
dir()
data<-read.csv(“Regression-Analysis-House Pricing.csv“,na.strings = “”)
View(data)
```

OceanofPDF.com

Observation	Dist_Taxi	Dist_Market	Dist_Hospital	Carpet	Builtup	Parking	City_Category	Rainfall	House_Price	carpet_area
1	9796	5250	10703	1659	1961	Open	CAT B	530	6649000	1659
2	8294	8186	12694	1461	1752	Not Provided	CAT B	210	3982000	1461
3	11001	14399	16991	1340	1609	Not Provided	CAT A	720	5401000	1340
4	8301	11188	12289	1451	1748	Covered	CAT B	620	5373000	1451
5	10510	12629	13921	1770	2111	Not Provided	CAT B	450	4662000	1770
6	6665	5142	9972	1442	1733	Open	CAT B	760	4526000	1442
7	13153	11869	17811	1542	1858	No Parking	CAT A	1030	7224000	1542
8	5882	9948	13315	1261	1507	Open	CAT C	1020	3772000	1261
9	7495	11589	13370	1090	1321	Not Provided	CAT B	680	4631000	1090
10	8233	7067	11400	1030	1235	Open	CAT C	1130	4415000	1030
11	4278	10646	8243	1187	1439	Covered	CAT A	1090	7128000	1187
12	8066	11149	12936	1751	2098	No Parking	CAT B	720	5762000	1751
13	7693	9130	14684	1746	2064	Open	CAT B	1050	6047000	1746
14	5236	10853	13054	1615	1931	Covered	CAT B	1160	5913000	1615
15	6027	6707	10176	1469	1756	Open	CAT B	770	6636000	1469
16	9648	14789	12812	1644	1950	Covered	CAT A	790	7887000	1644
17	11079	13102	13076	1578	1907	Open	CAT A	1440	7725000	1578
18	6698	11519	13441	1703	2045	Covered	CAT B	670	3817000	1703
19	9609	9066	13304	1438	1731	Open	CAT A	1030	6354000	1438
20	6209	7839	10660	1837	NA	Open	CAT B	790	4922000	1837
21	8155	8085	9837	1940	2340	Covered	CAT C	980	4019000	1940
22	9669	12385	13589	1421	1700	No Parking	CAT C	370	4346000	1421

When you enter the data into R, you need to check the class of the frame in your data.

class(data)

When you do this, you will obtain the output shown below. In this output, you can see that the dataset is stored in R in the form of a data frame.

[1] “data.frame”

We need to perform the next step to check the number of columns or rows present in the data frame. The following is the code to use.

dim(data)

The output is: [1] 932 10

From the above output, we see that the data set has 10 columns and 923 rows. You can use the following function to view the columns and rows' details in the data set.

summary(data)

This leaves you with the following output:

OceanofPDF.com

```
> summary(data)
  Observation      Dist_Taxi      Dist_Market    Dist_Hospital       Carpet
Min.   :  1.0   Min.   :  146   Min.   : 1666   Min.   : 3227   Min.   :  775
1st Qu.:233.8   1st Qu.: 6476   1st Qu.: 9354   1st Qu.:11302   1st Qu.: 1318
Median :466.5   Median : 8230   Median :11161   Median :13163   Median : 1480
Mean   :466.5   Mean   : 8230   Mean   :11019   Mean   :13072   Mean   : 1512
3rd Qu.:699.2   3rd Qu.: 9937   3rd Qu.:12670   3rd Qu.:14817   3rd Qu.: 1655
Max.   :932.0   Max.   :20662   Max.   :20945   Max.   :23294   Max.   :24300
                NA's   :13      NA's   :13      NA's   :1       NA's   :8
    Builtup              Parking    City_Category    Rainfall       House_Price
Min.   :  932   Covered      :188   CAT A:329     Min.   :-110.0   Min.   :    30000
1st Qu.: 1583   No Parking   :145   CAT B:365     1st Qu.: 600.0   1st Qu.:  4658000
Median : 1774   Not Provided:227   CAT C:238     Median : 780.0   Median :  5866000
Mean   : 1795   Open         :372                 Mean   : 785.6   Mean   :  6084695
3rd Qu.: 1982                                    3rd Qu.: 970.0   3rd Qu.:  7187250
Max.   :12730                                    Max.   :1560.0   Max.   :150000000
NA's   :15
```

Step Two: Visual Exploratory Analysis

After you import the data into R, you can use visual analysis to learn more about the hidden parameters. You can use the following plots to look at the data visually:

- Histogram
- Box Plot

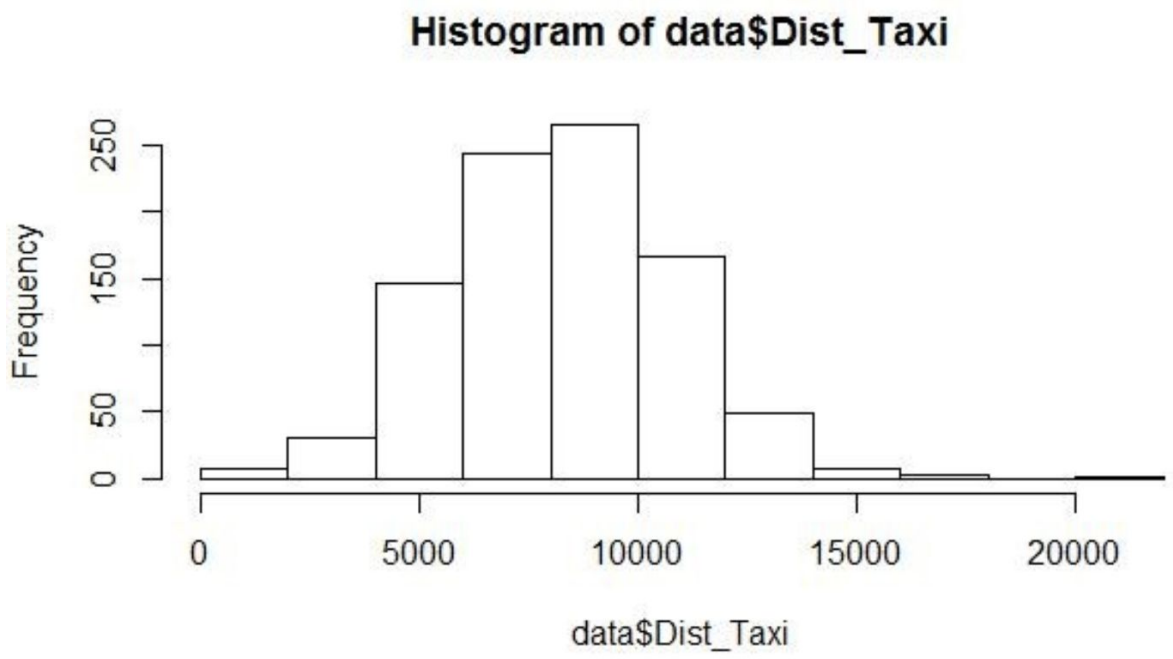
Histogram

A histogram is a visualization tool used to look at the data set on the basis on its numerals. Using a histogram, it becomes easy to determine if the data is distributed using a unimodal and bi-modal distribution. Most data sets are normally distributed, so if your data is not distributed normally, you can convert the data into a normal distribution. A histogram can also help you determine if there are any outliers in the data set or the column under study. Use the code given below to plot the data as a histogram.

install.packages(“plyr”)
library(plyr)

OceanofPDF.com

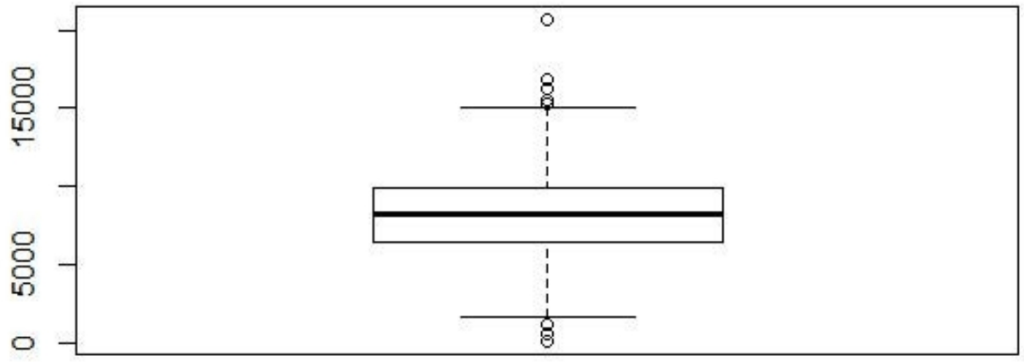
hist(data$Dist_Taxi)



Box Plot

A box plot is extremely useful since it shows you how skewed the data is. A box plot shows you the median and quartiles in the data set. You can also use this to spot any outliers in the data set. Use the code below if you want to visualize the data in the form of a box plot:

boxplot(data$Dist_Taxi)



OceanofPDF.com

Step Three: Correct Errors

In this step, you should identify the errors in the data set, and determine methods and models you can use to correct the errors in the code. You can also change names of the data frames to differentiate between the accurate frames used. The syntax to change the name of a data frame is:

data$carpet_area<-data$Carpet

Using the above syntax, you can rename the column 'Carpet' to carpet_area. There may be times when the data set imported in the language are stored as numeric values. In these cases, you need to change the column type using the following code below:

data$Dist_Taxi<-as.character(data$Dist_Taxi)
class(data$Dist_Taxi)
"character"

You can perform numerous array conversion operations in R. Some of these operations are listed below:

as.character()
as.numeric()
as.integer()
as.logical()
as.factor()

Since most data collected by organizations is unstructured or semi-structured,

OceanofPDF.com

you need to find ways to transform data as needed. String manipulation comes in handy since most elements are text based. You can change the text from lowercase to uppercase and vice versa in a specific column. To do this, you can perform the operation below:

```
#Making all uppercase
data$Parking<-toupper(data$Parking)
#Making all Lowercase
data$Parking<-toupper(data$Parking)
```

You can also remove any whitespaces or additional spaces within the data using the syntax below. You can do this to specific cells or columns, as well.

```
#Installing and loading the required packages
install.packages("stringr")
library(stringr)
#Trimming all whitespaces
data$Dist_Taxi<-str_trim(data$Dist_Taxi)
```

If you want to replace a specific letter or word in a specific column, use the code below:

```
#Replacing "Not Provided" with "Not Available"
data$Parking<-str_replace(data$Parking,”Not Provided”,”Not Available”)
```

If you want to replace any outliers in the data set using some summary statistics, like the median or mean, you can use the code below:

OceanofPDF.com

```
#Replacing the outliers of a particular column with median
vec1<-boxplot.stats(data$Dist_Taxi)$out;
data$Dist_Taxi[data$Dist_Taxi %in% vec1]<-median(data$Dist_Taxi)
```

The next steps will help you determine how you can identify and remove any missing values in the data set:

```
#Checking for missing values in the entire dataframe
any(is.na(data))
#Checking for the total number of missing values in the entire dataframe
sum(is.na(data))
#Checking for the total number of missing values in a particular column
sum(is.na(data$Dist_Taxi))
#Eliminating missing values completely from the entire dataframe
na.omit(data)
#Eliminating missing values completely from a particular column
na.omit(data$Dist_Taxi)
#Replacing the NA's in the entire dataframe with '0’s
data[is.na(data)]<- 0
#Replacing the NA's in a particular column with '0’s
data$Dist_Taxi[is.na(data$Dist_Taxi)]<-0
#Replacing the NA’s in a particular column with a summary statistics like median
data$Dist_Taxi[is.na(data$Dist_Taxi)]<-median(data$Dist_Taxi)
```

Use the tools below if you want to combine two or more columns in the data set:

OceanofPDF.com

```
#Installing and loading the required package
install.packages("tidyr")
library(tidyr)
data1<-unite(data = data,col = city_category_with_parking,City_Category,Parking)
View(data1)
```

The function 'unite' uses the following arguments:

- Data frame
- New column name
- First column
- Second column

The last two variables in the code are the names of the columns you want to combine.

You can also separate columns if needed:

```
data2<-separate(data = data1,city_category_with_parking,c("City_Category","Parking"), sep = "-")
View(data2)
```

The separate function has four arguments:

- Data frame
- Columns to separate
- Name of the new column
- Indicator used to separate the data in the column

When you follow the steps given above, you will end up with a clean data set.

OceanofPDF.com

Chapter Ten: Develop a Predictive Model Using R

If you want to understand what predictive modeling is and the strategic areas, let us break the process down into the essential components. You can divide the process into four parts. Every component needs a certain amount of time to execute the process. Let us look at each step and consider the time it will take:

1. Performing descriptive analysis on the data – this takes 50% time
2. The next step is to treat the data and identify any outliers and missing values – 40% time
3. Data modeling – 4% time
4. The estimation of performance – 6% time

These are only assumed times since the time taken by each programmer will vary depending on how well you know the concepts and language. Using the above data, you can cut the time down. Now, let us put the steps discussed above into action. The code used below is not the entire code, but it gives you an idea of what type of code to write. This is a skeleton of the entire algorithm in R.

The first thing to do is to split the entire data set into training and testing data set. Once you do this, you need to import the training data set into R. Use the code below to do this:

```
setwd("C:\\Users\\Tavish\\Desktop\\Kagg\\AV")
complete <- read.csv("complete_data.csv", stringsAsFactors = TRUE)
```

OceanofPDF.com

The next step is to read the names of the columns and the summary in the data set.

```
colnames(complete )
[1] "ID" "Gender" "City"  "Monthly_Income" "Disbursed" "train"
```

When you extract the summary, you should look for the following data:

- Variables
- Numeric variables
- Factor variables or predictors
- Target variables

If you notice any values missing in the data set, you need to create flags. Use the example below:

```
missing_val_var <- function(data,variable,new_var_name) {
data$new_var_name <- ifelse(is.na(variable),1,0))
return(data$new_var_name)}
```

Now, impute the values in the data set:

```
numeric_impute <- function(data,variable) {
mean1 <- mean(data$variable)
data$variable <- ifelse(is.na(data$variable),mean1,data$variable)
return(new_var_name)
}
```

You can also input other categorical variables in the data set to ensure every missing value in the code says 'Null.' The next thing to do is to pass any imputed variables in the process.

```
create_model <- function(trainData,target) {
set.seed(120)
```

OceanofPDF.com

```
myglm <- glm(target ~ . , data=trainData, family = "binomial")
return(myglm) }
```

The final step is to make the necessary predictions.

```
score <- predict(myglm, newdata = testData, type = "response")
score_train <- predict(myglm, newdata = complete, type = "response")
```

Once you develop this model, you can test the model using the testing data set.

```
auc(complete$Disbursed,score_train)
```

OceanofPDF.com

Chapter Eleven: Create Data Visualization to Communicate the Insight

Once you complete your analysis of the data, you need to use visualization techniques to explain to the stakeholders what you found in the data set. Visualization patterns show people the insights and patterns in the data set, which cannot be absorbed or shown in any other way. It is important to note that data is just bits and pieces of information, but you need to visualize it if you want to make sense of the information available.

In this chapter, we look at what data visualization is in a broad aspect. For instance, some people may look at loading data into excel or any other visualization platform as a form of data visualization. You can then convert the data imported into the tool into a picture of the screen. Having said that, you cannot learn everything about the data set using a table because tables do not help you identify any patterns, hidden or specific, in the data set immediately. A common example of this is the geographical pattern that is studied on a map only after you visualize it.

Bear in mind that every data visualization tool or software allows you to identify a hidden story or pattern in the data set. When you try to visualize the data using different tools, you learn about new patterns in the data set. You may have identified some patterns and relationships in the data set when you had initially downloaded it.

OceanofPDF.com

There may be other inferences that you may not have identified. Some insights may be due to an error in the data set, while others may result in the development of new insights. You can only identify these patterns when you visualize the data.

You can obtain a new direction and perspective of the data points and variables in the data set through data visualization. You can visualize data in numerous ways. If you have very small data points in the data set, you can easily identify the patterns using a table. A table has the labels, and the values in the data set organized in a structured manner. You can improve the analysis and understanding of the data set by filtering and sorting the data. Tables, however, have limitations. You can use tables to identify and understand the data set, and it’s easy to use tables to compare and understand the difference in the data points. If you want to compare multi-dimensional data sets, do not use tables.

You can also use charts to map the various data points and variables in the data set. These charts allow you to plot the data points in geometric shapes. You can also use scatterplots and other graphs to plot the data. Scatterplots can be used to map two-dimensional data sets. You can also use colors and thermal imaging to learn more about the data points in the set.
Most data collected by organizations are related to people. For example,

- Businesses may collect information about customers to understand and analyze their purchase history and preferences
- They may also collect historical data to determine a customer's credit history

OceanofPDF.com

- Businesses can collect information about financial transactions to determine when a customer may spend more or less money
- Governments and other legislative bodies may collect information about the crimes in certain areas and determine which region may have more crimes when compared to others. They may also collect information to determine how some areas in the city or country are doing when compared to others.

One of the most common forms of visualization is graphs. A graph allows you to identify the interconnections in the data sets. You can calculate each data point's position by using graph layout algorithms, and these algorithms make it easier for you to identify the structure in the data set easily. You need to find a way to model the network, and this is an important aspect to consider when it comes to data visualization. Bear in mind that every data set will have some obvious and hidden relationships.

Once you visualize the data using different visualization techniques, you need to understand and find any information from the data set. When you use visualization techniques, consider the following questions:

- What can you see in the image?
- Are there some patterns you have missed?
- Is this visualization exactly what you expected?
- What would happen to the data set if you included more data points?

OceanofPDF.com

You may end up with visualizations that do not reveal the information you are looking for. All you should do is try a new visualization. This way, you learn something new about the data.

OceanofPDF.com

Chapter Twelve: Document Your Insights from the Data

In most situations, you work with new data. You may develop some expectations or assumptions about the data set before you look at the data. It is best to make a note of these assumptions based on what you think about the data set. This will help you determine if you are biased towards specific trends in the data set. It also reduces the probability of misinterpreting the data set's variables or data points because of the original bias. This is an important tip to keep in mind when it comes to analyzing data.

It is for this reason documentation is extremely important in the process of data analysis. This is the step most people often skip. In the process of documentation, you need to plot the data points and understand them better. You may have created numerous charts as part of the visualization, but these can be confusing to understand. It is for this reason you need to take some time to document the following information:

- Why was this chart created?
- Has the data been updated to fit it into the chart?
- What information can I gather from the chart?

The information you document will help the stakeholders understand what you are trying to convey through the visualization.
Let us now look at which tool you can use to visualize the data in your analysis. When you look at the tools separately, you identify what each tool is good at.

OceanofPDF.com

It makes sense for you to consider tools that allow you to take data wrangling and visualization into account. You can choose to use different tools if you want to perform the processes separately. When you separate tasks, you need to import and export large volumes or chunks of data. The following are some tools:

- Non-programming visualization tools like Tableau
- Visualization libraries, for example, d3.js, Flare
- Spreadsheets like Excel or Google Docs.
- A programming framework like R, Python, or Pandas

Data Wrangling software like Google refine or Datawrangler

OceanofPDF.com

Conclusion

Data science is a new field of science that is used by organizations to analyze and interpret large volumes of data. This analysis helps businesses make informed decisions about the business. If you are new to data science, this book has all the information you need about data science.

The book introduces the concept of data science and the various processes needed to perform data analysis. You will learn about the various skills, both technical and non-technical, you need to develop to become a data scientist. The book also sheds some light on what different algorithms you can use to become a data scientist.

This book has some information about the different applications of data science. You can learn to implement different data science tools and techniques you learn in this book in different areas of your life. It is important to note that the skills mentioned in the book are essential to develop if you want to become a data scientist. What is most important to understand is that data science is a new field; you need to develop the skills if you want to identify any hidden patterns in the data set.

Bear in mind that you cannot learn everything at the same time. You need to give yourself some time to understand the basics. Once you do this, you can work on improving those skills to help you become an indispensable part of an organization. You can help businesses make informed decisions.
Thank you for purchasing the book. I hope you got the information you are looking for.

OceanofPDF.com

References

Ahuja, R. (n.d.). Introduction to Data Science. Coursera website: https://www.coursera.org/specializations/introduction-data-science

Castrounis, A. (n.d.). What Is Data Science, and What Does a Data Scientist Do? InnoArchiTech website: https://www.innoarchitech.com/blog/what-is-data-science-does-data-scientist-do

Data Science Tutorial for Beginners: What is, Basics & Process. (n.d.). www.guru99.com

Holtz, D. (2019, March 27). What is Data Science? 8 Skills That Will Get You Hired in Data | Udacity. Udacity website: https://blog.udacity.com/2014/11/data-science-job-skills.html

Martin, R. (2019, June 10). Linear Regression for Predictive Modeling in R –.Dataquest website: https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/

Perfect way to build a Predictive Model in less than 10 minutes. (2019, June 11). Analytics Vidhya website: https://www.analyticsvidhya.com/blog/2015/09/perfect-build-predictive-model-10-minutes/

Pros and Cons of Data Science - Why Choose Data Science for Your Career - DataFlair. (2019, March 27). DataFlair website: https://data-flair.training/blogs/pros-and-cons-of-data-science/

Sharma, H. (2017, January 5). What Is Data Science? A Beginner's Guide To Data Science. Edureka website: https://www.edureka.co/blog/what-is-data-science/

Sharma, M. (2018, May 2). DATA CLEANING USING R. Data Analytics Edge website: http://dataanalyticsedge.com/2018/05/02/data-cleaning-using-r

OceanofPDF.com